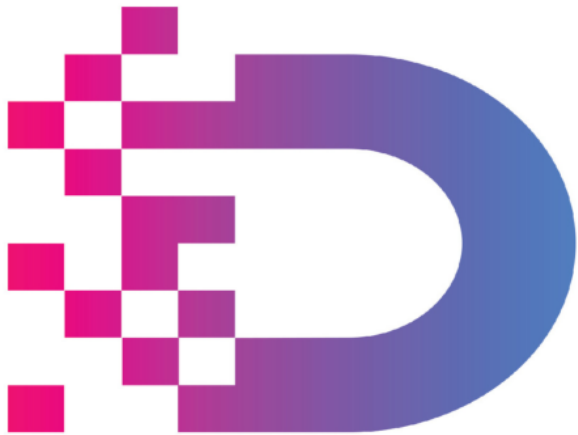
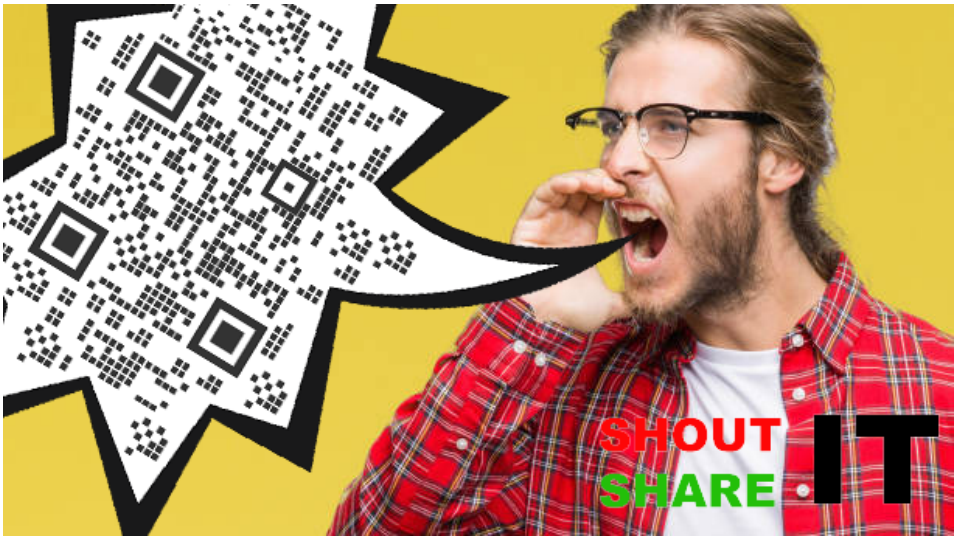
АЛЕКСЕЙ БЛАГИРЕВ



BIG DATA

ПРОСТЫМ ЯЗЫКОМ


SHOUT
SHARE
IT

Annotation

Наш телефон знает о нас больше, чем мы думаем. Он умеет собирать и анализировать информацию о том, как мы передвигаемся по городу, какие посты лайкаем и какими приложениями пользуемся. Он сообщит о пробках и поторопит на работу, чтобы мы не опоздали; подберет музыку под наше настроение и составит список персональных рекомендаций, чем можно занять себя в течение дня. Телефон – больше не устройство, по которому звонят, это уже средство управления окружающим нас миром. Незаметно мы окружили себя такими интерфейсами, которые создают невидимый барьер между человеком и окружающей средой. Планирование, управление, коммуникация, все теперь строится через эти программы и девайсы. Даже человеческие отношения. Но насколько глубока кроличья нора? Каждому предстоит разобраться в этом самому. Эта книга поможет донести основные принципы проектирования и создания таких интерфейсов управления бизнесом, обществом и окружающим нас миром посредством Больших данных. Читайте, наслаждайтесь и помните: сожжение книг противозаконно.

- Алексей Благирев
  -
  - Предисловие
  - Глава 1
    - Марсианские диалекты
    - Что же это все-таки такое и откуда взялось?
    - Постинформационное общество[4]
    - Data-driven организации[11]
    - 7 шагов data-driven decision culture
    - В чем ценность data-driven организации
    - Data-informed организации
    - Data-informed или data-driven
    - Революция open-source и доступность технологий
    - 4-я промышленная революция, или Почему человек больше не нужен для поиска инсайтов
  - Глава 2
    - С чего начинается стратегия данных?
    - Ключевые стейкхолдеры
    - Техническая инфраструктура
    - Зачем нужна стратегия данных?

        - Как влияет культура компании на успешность стратегии?
        - Кто владелец стратегии данных?
        - Self-service BI
        - Как измерить успешность стратегии данных?
        - Сколько стоит реализовать стратегию данных?
    - Глава 3
        -
        - Идеальная история отвечаем на ключевые вопросы
        - Декодирование аналитического контента требует усилий
        - Impact investment – у каждого рассказа должна быть цель
    - Глава 4
        -
        - Суровые европейские консерваторы
    - Глава 5
    - Глава 6
        -
        - Основные методы управления качеством данных
        - Как измерять качество данных?
        - Как понять, какие измерения качества выбрать?
        - Инструменты управления качеством данных
    - Глава 7
        - PaaS и платформы
    - Глава 8
        -
        - Проблемы с Big Data сегодня
        - Мы думаем, что понимаем Big Data
        - Как рассчитать финансовый эффект?
        - Big Data может быть вообще не нужна
        - К чему мы движемся? Тренды
        - Машинное обучение применяется все чаще
        - Послесловие
- notes
    - 1
    - 2
    - 3
    - 4
    - 5
    - 6
    - 7
    - 8

- ○ 9
- ○ 10
- ○ 11
- ○ 12
- ○ 13
- ○ 14
- ○ 15
- ○ 16
- ○ 17
- ○ 18
- ○ 19
- ○ 20
- ○ 21
- ○ 22
- ○ 23
- ○ 24
- ○ 25
- ○ 26
- ○ 27
- ○ 28
- ○ 29
- ○ 30
- ○ 31
- ○ 32
- ○ 33
- ○ 34
- ○ 35
- ○ 36
- ○ 37
- ○ 38
- ○ 39
- ○ 40
- ○ 41
- ○ 42
- ○ 43
- ○ 44
- ○ 45
- ○ 46
- ○ 47
- ○ 48
- ○ 49
- ○ 50
- ○ 51
- ○ 52

- ○ 53
- ○ 54
- ○ 55
- ○ 56
- ○ 57
- ○ 58
- ○ 59
- ○ 60
- ○ 61
- ○ 62
- ○ 63
- ○ 64
- ○ 65
- ○ 66
- ○ 67
- ○ 68
- ○ 69
- ○ 70
- ○ 71
- ○ 72
- ○ 73
- ○ 74
- ○ 75
- ○ 76
- ○ 77
- ○ 78
- ○ 79
- ○ 80
- ○ 81
- ○ 82
- ○ 83
- ○ 84
- ○ 85
- ○ 86
- ○ 87
- ○ 88
- ○ 89
- ○ 90
- ○ 91
- ○ 92
- ○ 93
- ○ 94
- ○ 95
- ○ 96

- ○ 97
- ○ 98
- ○ 99
- ○ 100
- ○ 101
- ○ 102
- ○ 103
- ○ 104
- ○ 105
- ○ 106
- ○ 107
- ○ 108
- ○ 109
- ○ 110
- ○ 111
- ○ 112
- ○ 113
- ○ 114
- ○ 115
- ○ 116
- ○ 117
- ○ 118
- ○ 119
- ○ 120
- ○ 121
- ○ 122
- ○ 123
- ○ 124
- ○ 125
- ○ 126
- ○ 127
- ○ 128
- ○ 129
- ○ 130
- ○ 131
- ○ 132
- ○ 133
- ○ 134
- ○ 135
- ○ 136
- ○ 137
- ○ 138
- ○ 139
- ○ 140

- ○ 141
- ○ 142
- ○ 143
- ○ 144
- ○ 145
- ○ 146
- ○ 147
- ○ 148
- ○ 149
- ○ 150
- ○ 151
- ○ 152
- ○ 153
- ○ 154

Алексей Благирев
Big data простым языком

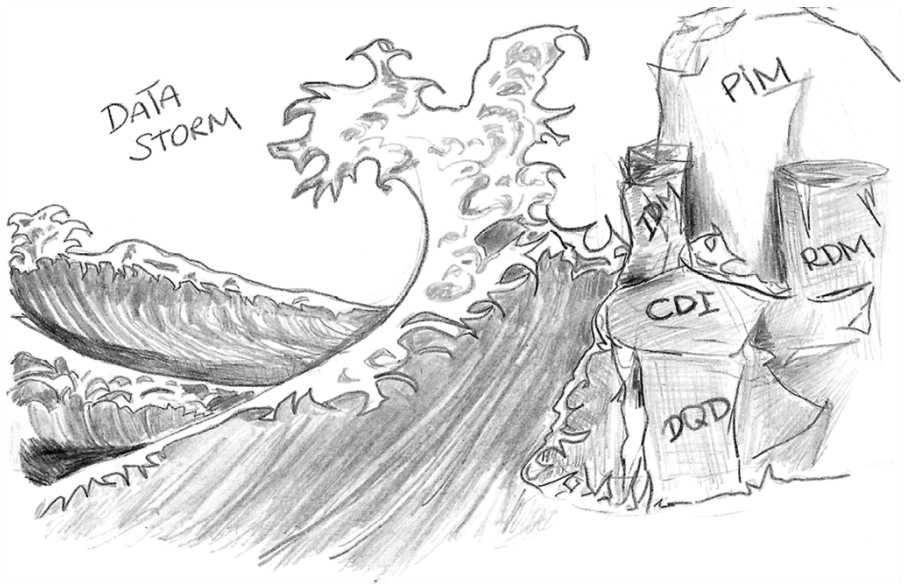
© Благирев А., текст, иллюстрации
© ООО «Издательство АСТ»

Предисловие

Люблю людей.

Именно такие мысли остаются в голове, когда тебе предлагают полностью переписать книгу. А если вы читаете это, значит, мне удалось, и я все-таки ее переписал.

Началось все с того, что один мой друг спросил, не знаю ли я людей, которые могли бы простым языком написать про Большие данные. Тогда я сразу представил бесконечное количество писем от издательства, разговоров, уточнений, переписываний, – всей этой суеты, и первое, что мне хотелось ответить: «Нет, таких разумных существ я не знаю».



Да и смысл писать про Большие данные, если про них уже столько всего написано и рассказано? Вероятность написать что-то умное – минимальна.

И вот я начал писать… Я даже уже представлял себе, как героически заканчиваю эту книгу, становлюсь миллионером и на все деньги с продаж бесконечных тиражей иду погашать ипотеку.

План был гениален, оставалось только его воплотить.

Но, когда я показал плод своих трудов редактору, он сказал, что книга сложна для восприятия, иными словами, подходит только для ботанов. Я честно писал ее с использованием книжной лексики, сложных эвфемизмов, деепричастных оборотов и кропотливо вставлял в текст ссылки на источники, если вдруг

упоминал материалы других авторов.

Один раз даже пришлось взять DMBOK, такую специальную «поваренную книгу» с инструкциями и стандартами организации работы с данными. Я перевел из нее целую главу на русский, но мне сказали, что это точно «слишком мощно» для читателя, как и попытка проанализировать существующее регулирование данных.

Итак, в поисках правды, баланса и закрытия личных гештальтов – как сейчас популярно говорить, мне дали книгу «Хулиномика» как пример образцовой книги жанра нон-фикшн.

Когда я взял в руки «Хулиномику», помимо ненормативной лексики в мыслях у меня появились смелые очертания нового эксперимента, поэтому вы держите в руках книгу про Большие данные, изданную под влиянием уникальной простоты и творческой логики изложения.

Мир данных – это компот, из которого трудно отделить то, что нужно знать, а что нет. И вроде бы все интересно, про все можно рассказать, но как понять, что из этого важно, например, учителю физкультуры, который на досуге решил погрузиться в данные?

Задачка оказалась сложнее, чем я думал.

Если вы пишите, скажем, про физику, то план изложения поправит научный редактор. А тут – технологии, англицизмы, и людей, знающих ключевые понятия, широту и многогранность Больших данных в издательстве просто-напросто нет.

И я взялся за дело. Сам.

Для начала я решил, что в каждой главе будет два уровня сложности. Первый – для тех, кто собрался почитать про данные, сидя на белом друге в тихой комнате, второй – для тех, чья сфера деятельности связана с данными.

Я написал большую главу про стратегию данных для тех, кто вынужден проектировать стратегию с нуля; попытался разобраться, как данные влияют на корпоративное управление компаниями; показал на ошибках людей, рисующих сложные, малопонятные графики, что формат изложения информации не менее важен, чем сам процесс получения знания.

Конечно, то, что вы держите в руках, – сильно переработанный вариант, но не менее достойный. Наверное.

Сегодня этот компот под названием «мир данных» – уже целая экономика, которая сильно повлияла на все вокруг, включая людей. Теперь нашими данными располагают голосовые помощники, а банки и компании, с которыми мы когда-либо имели дело, все чаще напоминают о себе и требуют внимания. Наш телефон знает, когда мы собираемся на работу, и заранее подгоняет нас к выходу, чтобы мы не опоздали из-за пробок, а

когда мы выбираем песню, которую хотим послушать в машине, он выдает нам подходящий плейлист.

Важно знать, что за данные, а точнее за искусственный интеллект, начали активно «топить» в обществе и бизнесе, поднимая проблемы этики их использования.

Просто задумайтесь, вся цифровая среда уже оперирует такими понятиями как «лайки», «репосты», «конверсии». Люди уже обсуждают, как и где подешевле купить трафик себе на сайт, а накруткой подписчиков в Инстаграме не пользуется только ленивый.

Мы оставили позади (в первой версии книги) весь романтизм и большие надежды, поместив в новую версию экспертное мнение по основным блокам работы с данными.

Читайте, наслаждайтесь и помните: сожжение книг противозаконно.

Алексей Благирев

Глава 1
Что такое Big Data?

Марсианские диалекты

О Больших данных, или Big Data сегодня знают все.

Или еще нет?

Регулярно данные обсуждаются на сложных конференциях, где популярные компании собирают под своими тентами от дождя пару тысяч молодых людей, размещают роботов и плюшевые пуфики, предлагают даже сыграть в игру с ботом, чтобы посетители могли поучаствовать в машинном обучении. Происходит это примерно так: за ограниченное количество ходов игроку необходимо как можно быстрее споить девушку-робота.

В общем, кто чем пытается покорить свою аудиторию, рассказывая о работе сервисов с данными. Вот только ни у кого нет единой картины.

Одни компании говорят про конфиденциальность, другие – про машинное обучение, перечислять можно бесконечно. Есть даже гипотеза о том, что общая картина больше никому не нужна.

«Как это не нужна?» – спросите вы и поспешите на ее поиски.

Выйдете вы из зоны комфорта, пройдетесь по ключевым конференциям, связанным с данными, прочтете статейки известных умных авторов, но все равно толком ничего не соберется вместе.

Чтобы погрузиться в эту тему, надо взять лопату и копать, копать, копать: по кусочкам собирать смыслы, общаться с разными людьми. Администраторы баз данных могут рассказать вам о том, как настраивать кластеры, а ребята, которые копаются в аналитике, помогут разобрать общую логику процесса.

Только вот почему-то каждый эксперт понимает один и тот же термин по-своему. Будто люди строили Вавилонскую башню из данных, чтобы достучаться до небес, а в конце концов все равно заговорили на разных языках, как написано в Ветхом завете. И эти эксперты вкладывают в, казалось бы, обычные слова, какое-то свое понимание, близкое только им.

Конечно, всех бы мог спасти робот-переводчик, который знает тридцать три наречия межпланетных иезуитов. Но, боюсь, пока его функционал не вырос до такого уровня, придется прикидываться оленеводами, которые впервые услышали о Больших данных. Надо признать, что в некоторых историях мне пришлось разбираться прям с самого что ни на есть нуля, так что расслабьтесь и получайте удовольствие. Будет весело!

А начнем с того, что познакомимся с народом.

#1

Есть такие важные и бессмертные инженеры по машинному обучению. Задача их проста – проектировать логику и обучать алгоритмы, известные как нейронные сети, заводя в них все новые и новые данные. Если спросить этих инженеров о чем-нибудь другом из области данных, то в большинстве случаев они понятия не будут иметь, о чем их спрашивают – например, кто такие дата-стюарды?

#2

Дата-стюарды и инженеры качества данных – это такие человечки, которые все правят, чинят и спасают, как Мастер Феликс-младший из игры Fix-It Felix Jr, по ней еще несколько лет назад сняли мультфильм «Ральф». Миссия стюардов и инженеров велика и необъятна. В данных всегда происходит переполох, и нужны те самые бравые ребята, которые прибегут со словами «я починю!». Они измеряют искажения в данных и исправляют те самые ошибки, которые допускают пользователи, работая с информацией.

Если спросить у них, в чем роль инженеров по машинному обучению и почему они вообще так называются, то, очень вероятно, что ответа мы не получим. И это нормально.

Разные бригады экспертов занимаются разной работой.

#3

Архитекторы и аналитики данных – это олицетворение разума. Они опираются на различные правила и методологию, чтобы структурировать данные внутри организации. Например, вместо обозначения таблички «N45» они напишут какое-нибудь гордое «Контрагент» и определят, что в этой табличке должна содержаться информация, касающаяся только контрагента, – например «ИМЯ» / «НАЗВАНИЕ», «ПАСПОРТ» / номер регистрации компании и так далее.

Суть архитекторов и аналитиков – стандартизировать взаимоотношения пользователей с данными и сделать самое главное: навести в этих данных порядок.

Результаты работы этих незаурядных личностей влияют через данные на управление организациями. По-умному их называют data-driven организациями. Они бывают разных типов и устроены

все по-разному, но описать data-driven организации или отличить их друг от друга сможет далеко не каждый из описанных специалистов. И это еще один большой вызов.

Разные профессии работы с данными разговаривают на разных языках и формируют собой организации нового типа, где люди не имеют единого представления о том, как ими управлять. Вопрос «чем отличается data-driven организация от data-informed организации?» введет в дичайший ступор не только читателя, но и экспертов, которые работают с данными каждый день.

Перспектива восприятия нового во многом касается наличия практических навыков. Конечно, сегодня мало кто из экспертов имеет руководящий опыт и был тем самым директором по данным, который пытался изменить мир, запуская трансформационные процессы в своей организации для того, чтобы повысить значение использования данных. Это прерогатива людей, которые стоят у руля, а они обычно не разбираются в технике, считая, что она не влияет на принимаемые с точки зрения развития бизнеса решения.

А это все не так. Свойства информационной среды, которые заложены в ней при ее проектировании, оказывают непосредственное влияние на объем и качество принимаемых решений в этой среде.

Когда люди учатся писать на таком языке программирования как Python, им не рассказывают, какие фреймворки проектирования хранилища данных существуют, и что работает, а что уже устарело. Не важно, откуда специалист, интересует его бизнес или IT, картина везде одна.

Получается, что знание сегментировано, утрировано и преподносится как тайное сокровище, хотя это не так.

Даже разработка на Python проста и похожа на обыкновенную разработку макросов в Excel.

Разбирая управленческие вопросы в организации, в части управления данными, стоит отметить самое важное и, наверное, самое главное. Гештальт, где должно определиться место функции управления данными или так называемого «директора по данным», до сих пор не закрыт и полон споров и противоречий.

IT-сфера активно определяет себя как поставщика данных и, соответственно, хочет играть в них ключевую роль, хотя большинство директоров в IT-сфере понятия не имеют, как правильно проектировать хранилища данных или функцию управления ими. Все ждут постановки от бизнес-подразделений.

Но сейчас ситуация, конечно, намного лучше, чем несколько лет назад, когда бюджеты заливались в бессмысленные проекты,

обреченные на смерть еще в пубертатном периоде использования технологии. Тогда пожилые дядечки в возрасте, которые рулили IT-департаментами, с большой долей вероятности были поклонниками Билла Инмона (автора первой книги по созданию хранилища данных) или Ральфа Кимбалла (антагониста Билла). Конечно, согласия между этими концептами мало, и все споры всегда превращаются в дедовские войны на лазерных мечах. Причем, у них разное мнение даже на счет того, как и какими инструментами правильно обрабатывать данные в этих хранилищах.

Например, основной подход – это обрабатывать данные по расписанию, используя специальные инструменты – программы (ETL или ELT) для этой задачи.

Современные эксперты запустили уже свою собственную религию о том, как правильно использовать данные и собирать их в специальную штуку под названием Data Lake. Некоторые из этих экспертов пошли так далеко, что даже отказались от привычных инструментов обработки данных (ETL или ELT), заменив их малопонятной парадигмой, – разбивая все алгоритмы обработки на одинаковые шаги и превращая эти шаги в отдельные программы (сервисы) для создания сложных алгоритмов обработки данных.

Я вам скажу так: все, что можно было когда-либо сделать в Больших данных и машинном обучении – уже сделано. Теперь нужно просто брать существующие методы и сервисы и показывать им новые данные, обучая тем самым алгоритмы

адаптироваться.

Перевожу на отечественный. Все, что осталось большинству специалистов – это участвовать в решении только одной задачи, загружать все больше данных для обучения уже существующих алгоритмов. Так ли это? Еще разберемся. Но такие мировые компании как Gartner, уже признают, что роль человека в кооперации с искусственным интеллектом отходит на задний план: необходимо предоставить искусственному интеллекту возможность учиться решать ежедневные задачи. Называется этот подход Augmented Intelligence.

В этой книге вместе представлены различные подходы и методы, которые в совокупности с заумной точкой зрения ведут читателя по новым путям работы с данными. Разобщенность терминологии и понятий, собственно, и подтолкнула меня к идее описать практический опыт тех решений, которые можно использовать для получения практического результата. Это должно помочь определить и выявить новые перспективы в работе с данными, чтобы освоить те дальние рубежи экономики, куда еще не проникла цифровизация.

Что же это все-таки такое и откуда взялось?

Начну со сложного. Понятие Big Data – это такое облако тегов, которое имеет несколько измерений, то есть зависит от ракурса, с которого смотрят.

Пространство координат, благодаря которым можно легко разобраться в том, что такое Big Data, постоянно меняется, создавая отдельные группы понятий, практически не связанных друг с другом. Трудно представить, да?

В Интернете есть известный мем о том, что в одном сперматозоиде содержится 37,5 мегабайт информации **ДНК**[1][2]. А в результате генерального «салюта» выдается порядка 1500 терабайт.

К слову, в 2013 году мне удалось стать участником крупнейшего внедрения в банковском секторе размером в 51 терабайт. Я внедрял хранилище данных Vertica от Hewlett-Packard. Когда моя команда поместила все транзакции одного крупного банка в это хранилище, у нас получилось немногим больше десяти терабайт. А тут почти в 30 раз больше. В 30!

Так что самые «большие» данные еще впереди.

А теперь просто. Понятие Big Data можно сравнить с термином «инди-рок», который появился в 80-х годах. Так называли стиль, напоминающий гаражный рок или брит-поп, который играли группы в колледжах или университетах. Благодаря журналистам этот термин обрел множество значений, трактовок и представлений, поэтому инди-роком все стали называть любой стиль музыки, который хотя бы издалека напоминал Oasis, Blur и другие подобные группы.

К чему это? Любую активность, которую я считаю хоть как-то связанной с жизненным циклом данных, я называю Big Data.

Когда понятие попадает в мейнстрим, оно становится #хэштегом, который позволяет привлекать общественное внимание. Да всем плевать на смысл этого хэштега, главное – чтобы было прикольно.

Это происходит, например, потому, что большинство журналистов и копирайтеров не понимают, с чем они столкнулись, что это за технология, и как она будет вести себя дальше. Никого особо не парит, как ее назвать.

Прямо как в издательствах. Это ведь прикольно. Ты пишешь книгу, а ее вычитывает редактор, который не понимает, что это такое.

Однажды мне рассказали историю о том, как один высокопоставленный чиновник участвовал в реализации законопроекта в области платежей, но сам при этом ни разу в жизни не сделал ни одного банковского перевода. С Big Data так же.

Лет десять назад термин Big Data воспринимался исключительно как инфраструктурный – под ним понимался специальный класс баз данных, которые позволяли быстро обрабатывать большие объемы информации. То есть, Big Data называлась просто категория железок (серверов), которые умели выполнять определенные вычисления.

Зачем они были нужны? Затем, что обычные железки не умели работать с большим количеством записей. Им было сложно. Памяти не хватало, процессоры грелись, пыхтели бедняги, а скорость расчетов оставляла желать лучшего. Железяки или сервера категории Big Data позволяли решить эту проблему. Потом придумали, что дело вовсе не в железяках, и что можно создавать программное обеспечение («софт»), которое будет работать на самых обычных настольных компьютерах, объединенных в единые узлы. Такие конструкции могли работать параллельно над конкретной задачей из области обработки данных. По-научному их называли «программными комплексами» и «кластерами».

Аудиофайлы, изображения, сложные и слабоструктурированные файлики в то время мало обрабатывались. Существовало сильное ограничение по их исследованию. Для них также требовалось специальное программное обеспечение, а у обычных баз данных не было возможности быстро провести анализ.

Технологии очень быстро эволюционировали. В какой-то момент на смену традиционному понятию Big Data пришел еще один новый термин – Smart Data. Он означал, что «Умные данные» – это сигнал, а «Большие данные» – шум. Таким образом появилась парадигма, разделяющая методы анализа: исследования «шумов» и выявления «сигналов».

За какие-то двадцать лет мир тряхануло так, что он перешел от рынка, где нельзя было купить данные интернет-трафика со «следами», оставленными пользователями, к рынку, где любые данные можно достать в любой момент.

И все бы ничего, но мир перевернулся с ног на голову. Данных стало так много, что их внезапно начали регулировать. Беспощадно и беспристрастно.

Одним из первых пострадал банковский сектор. Все процессы и продукты пришлось пересматривать, потому что теперь даже

для кредитного решения банк не может купить данные у кредитного бюро, чтобы проверить потенциального заемщика без его согласия.

В 2018 году появилось регулирование GDPR в Европе. Оно стало настолько жестким, что банки вынуждены были остановить привычные процессы привлечения клиентов в Интернете.

Конечно, если смотреть на все со стороны, то трансформация, которая произошла, – колоссальна. Только представьте, раньше данными занимались где-то внутри IT, в специальных операционных хранилищах (еще они называются ODS), дешево и сердито эти данные сваливались в одну кучу из разных источников. Но теперь мир перешел на новую стадию, данные – это новая нефть, из данных начали строить большой бизнес.

Новые технологии неизбежно приведут человечество к изменению мышления. Об этом уже писали эксперты[3], анализирующие влияние изучения другого языка на мышление человека. Новые технологии – это еще и переход к новой терминологии, который повлечет за собой новую форму организации взаимодействия потребителей и компаний. А она еще не выработана. Это значит, что данные как актив еще не имеют своей утвержденной и принятой формы по ведению бизнеса.

Поэтому теперь термин Big Data, скорее, отражает новую модель зрелости бизнеса, общества и государства, он больше не является просто олицетворением технологий хранения данных. Сегодня Big Data подразумевает, что пользователь понимает, как быстро и легально обработать информацию, и как ее структурировать таким образом, чтобы результаты этой работы были понятны окружающим.

Постинформационное общество [4]

Взрывной рост технологий использования данных приблизил человечество к новой модели своей работы – постинформационному обществу.

Звучит слишком заумно? Вообще префикс «пост» уже много где используется: постистория, постмодернизм, постиндустриальное общество и так далее.

Смысл постинформационного общества в том, что полезные знания среди разнообразной информации теперь могут находить алгоритмы, а не люди, которые их спроектировали.

Ну, то есть, учась в школе, ребенок может решать домашнюю работу вместе с алгоритмами, а не с родителями.

А еще с алгоритмами можно анализировать диагнозы множества пациентов или симптомов одновременно, не полагаясь на человеческую экспертизу.

Это реально?

Ага. Google со своим умным «движком» TensorFlow или Яндекс с CatBoost сделали возможным создание уникальных сервисов с использованием данных в домашних условиях (без каких-либо специальных лабораторий).

И чем больше мы используем алгоритмы, тем больше они учатся. Это можно гордо назвать демократизацией – когда всем понемногу достается кусочек счастья.

Демократизация технологий запустила новые процессы по унификации роли человека в процессах обработки, управления данными и развития искусственного интеллекта. Ручной труд стал больше не нужен. Всякие сверки и контроли – работа, которую теперь можно поручать алгоритмам, и они, в отличие от человека, умеют справляться с ней без ошибок.

Даже последний рубеж, которые машины взять никак не могли – тоже покорился. За несколько лет алгоритмы смогли освоить решение ранее сложных творческих и коллаборативных задач. Причем, этот рывок невозможно было спрогнозировать еще пять лет назад.

Такие системы как Alexa, Siri, Алиса и другие, ускоренными темпами захватывают рынок персональных ассистентов.

В 2015 году эксперты даже в своих самых смелых ожиданиях не могли сойтись в том, что алгоритмы смогут пройти этот рубеж всего лишь через год.

Сегодня есть ощущение, что близится еще один большой рывок, и он может произойти в ближайшие несколько лет.

По одной из гипотез им станет трансформация работы с

данными для производств. Тогда собираемая информация будет использоваться с целью анализа и выявления аномалий операционного цикла производства, упрощая управление конвейером, будь это надой молока с установленными датчиками на коровах или завод по производству металлической продукции. Я говорю о едином управлении жизненным циклом продукта или услуги, например – локомотива. Компании взаправду разрабатывают единую концепцию жизненного цикла локомотивов и цифровизации депо. Это уже происходит в России.

ПОЧЕМУ ЭТО ВАЖНО
ТРАДИЦИОННЫЙ БАНКИГ
ИНДУСТРИАЛЬНЫЙ БАНКИНГ
ДЛЯ БАНКА
• СЛОЖНЫЕ И ДОРОГИЕ ПРОЦЕССЫ
• УПРАВЛЕНИЕ СДЕЛКОЙ КАК ПРОЕКТОМ
• БОЛЬШИЕ РИСКИ
• СЛОЖНОСТЬ МОНИТОРИНГА ОБЪЕКТОВ ЗАЛОГА
• ЦИФРОВЫЕ И ОПТИМИЗИРОВАННЫЕ ПРОЦЕССЫ
• АНАЛИЗ РИСКОВ И ПОТРЕБНОСТЕЙ НА ОСНОВАНИИ ДАННЫХ ЧЕРЕЗ ЦИФРОВЫЕ КАНАЛЫ
• ДОСТУПНОСТЬ МОНИТОРИНГА СЛОЖНЫХ СДЕЛОК
ДЛЯ КЛИЕНТА
• МНОГО ДОКУМЕНТОВ ДЛЯ ЗАКЛЮЧЕНИЯ СДЕЛКИ
• МЕДЛЕННАЯ РЕАКЦИЯ БАНКА
• ВЫСОКАЯ СТОИМОСТЬ
• ПОЛУЧЕНИЕ ФИНАНСИРОВАНИЯ И КРЕДИТНЫХ ПРОДУКТОВ (ГАРАНТИЙ) БЕЗ ДОПОЛНИТЕЛЬНЫХ БУМАГ И СПРАВОК И БЕЗ ВЫЕЗДА СПЕЦИАЛИСТОВ
ДЛЯ РЫНКА
• НЕТ СИНЕРГИИ МЕЖДУ РАЗНЫМИ ФИНАНСОВЫМИ ПРОДУКТАМИ
• ВЫСОКИЕ БАРЬЕРЫ ДЛЯ ПОЛУЧЕНИЯ ФИНАНСОВЫХ ПРОДУКТОВ
• ЕДИНАЯ ЭКОСИСТЕМА И МАРКЕТПЛЕЙС ДЛЯ УЧАСТНИКОВ ФИНАНСОВОГО РЫНКА

ПОЭТАПНОЕ ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ БАНКИНГА
01 БАНК, СТРАХОВАЯ КОМПАНИЯ
ОТБОР ОТСЛЕЖИВАЕМЫХ ПОКАЗАТЕЛЕЙ ВЛИЯЮЩИХ НА ПРОИЗВОДСТВО
РАСЧЕТ НОРМЫ ПРОИЗВОДСТВА И ДОПУСТИМЫХ ОТКЛОНЕНИЙ ОТ НОРМЫ
ОЦЕНКА ВЛИЯНИЯ ОТКЛОНЕНИЯ ОТ НОРМЫ ПРОИЗВОДСТВА НА ДЕНЕЖНЫЙ ПОТОК ПРОЕКТА
ОЦЕНКА ВЛИЯНИЯ ДЕНЕЖНЫХ ПОТОКОВ ПО ПРОЕКТУ НА МОДЕЛЬ PD (ДЕФОЛТА) ЗАЕМЩИКА
02 ПРЕДПРИЯТИЕ
УСТАНОВКА ДАТЧИКОВ ДЛЯ МОНИТОРИНГА ПРОИЗВОДСТВЕННЫХ ПОКАЗАТЕЛЕЙ
ВКЛЮЧЕНИЕ ДАННЫХ В ЭКОСИСТЕМУ
SMART DATA
ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ ОТКЛОНЕНИЙ ОТ НОРМЫ ПРОИЗВОДСТВА НА ОСНОВЕ ДАННЫХ
ОПЕРАТИВНОЕ РЕАГИРОВАНИЕ (MARGIN CALL) НА УХУДШЕНИЕ ПАРАМЕТРОВ СДЕЛКИ И ПОМОЩЬ В ИХ РЕШЕНИИ НА ПРЕДПРИЯТИИ
УСТРАНЕНИЕ АНОМАЛИЙ, ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ПРОИЗВОДСТВА
03 ОБЪЕКТЫ ФИНТЕХА
СПЕЦТЕХНИКА И ДВИЖЕМОЕ ИМУЩЕСТВО
ЗАПАСЫ И ТОВАРНО-МАТЕРИАЛЬНЫЕ ЦЕННОСТИ
РАЗЛИЧНЫЕ ВИДЫ ПРОИЗВОДСТВА, В ТОМ ЧИСЛЕ CUSTOM ПРОИЗВОДСТВО
РАЗЛИЧНЫЕ ВИДЫ ПРОИЗВОДСТВА, В ТОМ ЧИСЛЕ CUSTOM ПРОИЗВОДСТВО

Создание подобных центров управления предприятиями сегодня не имеет технологических барьеров, проблема исключительно в кооперации участников. Решив ее, мир откроет невообразимую возможность создания адаптивной экономики, когда плановые значения заменяются на стандартные нормы производства, которые высчитывают алгоритмы в зависимости от множества факторов.

Но большинство людей все еще мыслит устаревшими категориями.

Для людей, проработавших много лет на производствах, все кажется достаточно понятным и простым. Сначала рисуешь и проектируешь с инженерами деталь, потом готовишь документацию, где прописываешь, как эту деталь обслуживать, потом производишь и, наконец, обслуживаешь.

ДАННЫЕ – ОСНОВА ИНДУСТРИАЛЬНОГО ФИНТЕХА
DIGITAL INDUSTRY
ИСПОЛЬЗУЮТ ДАННЫЕ ДЛЯ ЭФФЕКТИВНОГО УПРАВЛЕНИЯ ПРЕДПРИЯТИЕМ
SMART DATA
DIGITAL INSURANCE
ПРЕДОСТАВЛЯЮТ СТРАХОВЫЕ ПРОДУКТЫ НА ОСНОВАНИИ ДАННЫХ
DIGITAL CORPORATE BANK
ИСПОЛЬЗУЮТ ДАННЫЕ ДЛЯ РАСЧЕТА ПОТРЕБНОСТИ В ФИНАНСИРОВАНИИ И МОНИТОРИНГЕ ОБЪЕКТОВ ЗАЛОГА
ОСНОВНЫЕ ШАГИ
• ЗАПУСК АКСЕЛЛЕРАТОРА ДЛЯ ПОИСКА И ОТБОРА ПЕРСПЕКТИВНЫХ ПРОЕКТОВ НА БАЗЕ МБИ
• СОЗДАНИЕ ПЕСОЧНИЦЫ ПО ОБМЕНУ ДАННЫМИ НА БАЗЕ МБИ
• ПРОВЕДЕНИЕ ФОРУМА И СТРАТЕГИЧЕСКИХ СЕССИЙ НА БАЗЕ МБИ

Казалось бы, все цели ясны, все пути определены – вперед, товарищи!

А на деле все сложнее. Упомянутый выше локомотив может быть старой развалиной без документации. И вот тут людям приходится креативить. Иными словами, инженеры пытаются решить проблему на месте, прямо в депо. Таких примеров много. Что это означает? Только то, что привычного конвейера, который придумал в свое время Генри Форд, больше не существует. Признать это сложно.

Людям хочется верить, что всем можно управлять, а если запустить какой-нибудь сверхмасштабный проект, то вообще получится все вокруг цифровизировать и изменить. Потоки данных будут передаваться от производства к другим участникам рынка, например к страховой компании, которая будет выписывать страховой продукт, используя данные без выезда специалистов.

Страховая премия в этом случае может быть ниже рынка как минимум на десять процентов, при этом сам продукт будет более маржинален для страхового бизнеса, так как начнет использовать более точную оценку наступления риска, построенную на данных. Аналогичные продукты с использованием данных может предоставлять банковский сектор. Все соединится в единую экосистему обработки информации и извлечения из нее ценности.

Захватывает, не так ли?

Но на практике никакие масштабные программы не работают,

потому они медленные и не говорят на одном языке с технологией.

Государство, как и бизнес, тоже движется в сторону повышения роли данных. Но как сравнить, где находится государство с точки зрения роли Больших данных?

В 1965 году ООН ввел разделение на развивающие рынки и развитые страны.

Не важно, что это разделение уже не работает – его опроверг Ханс Рослинг[5]. Важно, что была попытка предложить систему оценки для сравнения экономического развития стран.

Сейчас, конечно, в национальном плане, единых критериев оценок до сих пор не выработано, хотя каждый малозначимый институт развития пытается предложить свою модель оценки для Больших данных. Короче, не понятно, кто где находится и куда идет.

Например, модель зрелости цифрового государства исследовательской компании Gartner, предполагает пять ступеней зрелости, где data-centric государство – это третья средняя ступень в развитии, этап, когда власть понимает, какие данные есть, когда она отладила процессы их получения и управления качеством.

Россия сегодня успешно завершает переход с первой ступени (E-Gov[6]) на вторую – когда для создания новых информационных сервисов федеральные и муниципальные органы власти предоставляют возможность получения открытых данных, хранящихся в государственных учреждениях. Но сами данные еще разрознены, некачественны, и, по сути, пользоваться ими пока что нельзя.

В Министерстве цифрового развития один из важнейших проектов – создание платформы управления классификаторами (для статистики), когда бизнес и общество могут стать основными источниками данных друг для друга. В идеале разработка платформы может устранить разобщенность классификации, например, номенклатуры товаров. Представьте себе, что больше не надо заполнять никакие накладные, таможенные декларации и прочие бумаги, весь товар регистрируется при производстве и отслеживается. Можно забыть про бумагу.

Единые классификаторы товарных позиций позволяют существенно упростить взаимодействие между несколькими торговыми рынками. В какой-то момент классификаторы позволят создать между ними уникальные зоны свободной торговли. Допустим, что вы приехали в аэропорт и идете через «зеленый коридор», вас никто не трогает, а рядом, в «красном коридоре», происходит принудительный досмотр вещей.

Мысленно вы улыбаетесь, радуетесь тому, что вас там нет. Представьте, что таким может быть производство, банкинг, страхование и торговля.

Помимо бизнеса или государства, конечно же, данные сами по себе точно так же оказывают непосредственное влияние на рядового пользователя, например, упрощая процедуру идентификации и получения тех или иных сервисов, в том числе и финансовых. Так, можно удаленно открыть банковский счет, используя только биометрические данные и информацию из учетной записи в государственных системах. Вот вам и опять какие-то новые интерфейсы, которые уже вроде как работают. Пора в них разобраться.

В мире давно существует множество платформ, таких как id.me, tupas, bank.id и других, позволяющих использовать единую учетную запись без необходимости хранить десятки паролей.

Эти платформы формируют будущее цифровой идентичности.

С другой стороны, новое общество, которое уже десятилетием пользуется социальными сетями, электронной почтой и мессенджерами, обзавелось уникальными артефактами и привычками, которых нет как в настоящей жизни, так и в юридической практике.

Такие понятия как «лайк», «репост», «шер», «трансляция», оказывают большее влияние на пользователей, чем пощечина. Из-за лайков люди ссорятся, расходятся или строят целые бизнес модели.

Появилось такое явление как «цифровое рабство» которое стало следствием того, что данные пользователей не принадлежат им самим. Во многом это помогло цифровым платформам проектироваться без учета общественного диалога. Но парадокс в том, что такой диалог был невозможен на момент зарождения таких платформ ввиду отсутствия пользовательского опыта по использованию данных у самого общества. Соответственно, нас будут ждать еще и этические дилеммы в отношении тех или иных данных.

Сегодня общество уже переживает рефлексию о том, что такое «хорошо» и что такое «плохо» в отношении своих данных. Что делать можно, а что делать не нужно, даже если это явно не запрещено законом, еще предстоит определить. Определение этой границы в использовании данных откроется в самое ближайшее время.

Необходимо разобраться в совершенно новых явлениях, в том числе таких, как «цифровое бессмертие».

Кто и как может пользоваться данными, если пользователя

больше нет среди живых? Стоит ли оставлять его «цифровые следы» во всемирной паутине?

В 2015 году в сети Facebook насчитывалось более пяти миллионов аккаунтов (страниц) людей, которые умерли. Вот вам так называемые цифровые памятники и реквиемы.

В нашей стране, если в табличку с персональными[7] данными добавить отдельное новое поле «дата смерти» и заполнить его, то такие данные перестанут быть персональными по действующему закону. Они более вне законодательного поля.

Или другой случай. Номер мобильного телефона сам по себе не является персональными данными и не защищен законом о персональных данных. Интересно, не правда ли?

С точки зрения регулирования в России сегодня есть базовый минимум по защите информации и прав пользователей, но как это законодательство реально работает по отношению к данным, предстоит открыть каждому индивидуально.

В развитом бизнесе количество информационных систем, хранящих данные пользователей, может исчисляться десятками, а порой и сотнями. Даже если в соответствии с законодательством рядовой пользователь напишет в компанию, которая обрабатывает его персональные данные, обращение с просьбой удалить их и реализовать свое законное право на «забвение»[8], компания не сможет полностью удалить или обезличить данные пользователя в своих системах. Во многом это произойдет из-за недостаточной развитости процессов и решений управления данными, ведь пользовательские данные в большинстве случае разбросаны по системам, не имеющим единой красной кнопки, которая бы могла все замаскировать.

Но существуют и исключения. Например, решение от команды HumanFactorLabs позволяет объединить и связать различные образы, копии и образцы данных о потребителе, клиенте или организации в разных системах, что в нужный момент позволит реализовать право на «забвение» посредством отдельного сервиса.

Регулирование данных в России представляет собой разобщенный процесс, по которому нет единого координирующего органа. Если мы вернемся к праву на «забвение», то существующая статистика обращений в России и за рубежом показывает, что большая часть исков пока проигрывается, да и количество обращений потребителей с исками невелико. Но переломный момент рано или поздно наступит.

Начиная с 2015 года[9] все налоговые агенты стали обязаны указывать в сопроводительных справках при удержании налога

информацию об идентификационном номере налогоплательщика (ИНН).

Для справки: организация является налоговым агентом, если платит за кого-то налоги как, скажем, работодатель, который платит за своего сотрудника налог на прибыль, или банк, который удерживает налог, если привлекает депозиты по высокой ставке или решил списать часть задолженности.

Ситуация усложнилась тем, что для финансовых организаций поле «ИНН» не являлось обязательным при выдаче банковского продукта (кредита или депозита). Добавление нового поля требовало организовать доработку всех ключевых банковских систем.

Непредоставление такой информации в Федеральную налоговую службу могло повлечь за собой возможность наложения штрафа на налогового агента в сумме от двухсот тысяч рублей за одну запись[10]. Сумма штрафа в пятьдесят миллионов рублей становилась существенной для ведения бизнеса с данными клиентов.

Таким образом, данные помимо возможности монетизации приводят организацию к риску получения внепланового расхода. Подход для работы с ними усложняется, требуются новые инструменты, новые профессии и новые правила работы с данными.

Данные – это актив, новая нефть, которая еще не имеет всех необходимых дефиниций по правильности или этичности использования.

Кругом только косяки и сложности. Чтобы понимать многообразие всех этих связей, которые появились, нужно обладать определенными навыками работы с данными как с точки зрения технологии, так и с точки зрения буквы закона.

Хочу упомянуть моего друга, Джозефа Маклеода. Он был когда-то UX дизайнером Nokia и является автором концепции Off-boarding. Согласно его парадигме, пользователи в цифровой среде ведут себя уже не так, как на индустриальном рынке. Они перестали бесконечно потреблять.

Информации вокруг стало так много, что внимание пользователей научилось чаще переключаться. Пользователям больше не нужно то, что им предлагали обычно. Капитализм в привычной форме отступает. Жизненный цикл потребителя теперь должен не только уверенно начинаться и поддерживаться, но и завершаться.

Завершение – один из важнейших этапов взаимодействия с пользователем в цифровом мире, но большинство компаний и

сервисов не уделяют ему должного внимания, из-за чего данные пользователей по-прежнему остаются в компаниях. Висят незакрытые банковские счета, приходят уведомления о подписках и сервисах, которые уже не интересны клиентам.

Data-driven организации[11]

Если вы работаете с данными, то необходимо помнить, что все новинки и важные изменения в подходах работы с ними всегда отражались в первую очередь в маркетинге или в коммуникации с клиентом, будь то UX-интерфейс или персональное уведомление.

В середине 2000-х, организации, занимающиеся дизайном, провозгласили новую тенденцию data-driven организаций, когда все расположения кнопочек, иконок или иных интерфейсов подчиняются логике работы на основании данных. Так называемый **Data-driven Design**[12].

Иными словами, все, что увеличивает конверсию, отражает текущее поведение клиента или потенциального клиента, должно строиться на основании данных и наблюдений. Получается, что все события превращаются в данные, которые ведут к конкретным решениям, так что организация становится дата-центрированной, то есть все решения внутри нее по созданию ценности, запусков продуктов или оптимизации, подчиняются исключительно данным.

Впервые термин data-driven был представлен в 1990 году Тимом Джонсоном[13], преподавателем School of English в Университете Бирмингема. Он предположил, что в основе любого языка находятся определенные общие понятия, «corpus»[14], на основе которых можно строить зависимость и исследовать лингвистику языка. Для своего исследования Джонсон использовал Международную базу лингвистических данных Бирмингемского университета (COBUILT). Эта работа легла в основу создания и описания корпусной лингвистики, что, в свою очередь, позднее повлечет за собой создание машиночитаемой лингвистики, использование Скрытых Марковских Моделей[15] и создание алгоритмов распознавания образов и текста.

Впоследствии централизация решений через данные распространилась на все ключевые бизнес-процессы без исключения и привела к новым формам внутренней работы организаций – data-driven organization.

Data-driven организации – это такие компании, в которых все внутренние процессы и большинство решений вокруг них строятся исключительно на основании данных. Вначале 2000-х ряд компаний провозгласили себя data-driven: Google, Facebook и другие.

Развитие новой формы кооперации человека с

использованием данных немедленно натолкнулось на один из первых барьеров на пути своего становления.

Им стал синдром HYPPO.

В 1963 году психолог из Йельского Университета, Стэнли Милгрэм, поставил эксперимент по социальной психологии, который описал позднее в статье «Подчинение: исследование поведения». Суть эксперимента сводилась к тому, что испытуемому предлагали стать на время Учителем и «помочь» Ученику (который был актером) выучить ряд слов и сочетаний. Экспериментатор дал Учителю указание, в случае ошибки, каждый раз бить током Ученика. При этом, каждая новая ошибка влекла за собой увеличение силы тока, вплоть до смертельно опасной. Ученик, в свою очередь, имитировал боль от тока, а Экспериментатор заставлял Учителя продолжать эксперимент, несмотря ни на какие возгласы со стороны Ученика.

До начала эксперимента Стэнли Милгрэм попросил большинство коллег, с которыми работал, оценить, сколько испытуемых дойдет до конца эксперимента. Большинство сошлось на двадцати процентах, но на практике все вышло ровным счетом наоборот. Менее двадцати процентов участников отказались продолжать эксперимент, а подавляющее большинство прошло его до конца. Этот психологический эксперимент показал чрезвычайно сильно выраженную готовность здоровых и нормальных взрослых людей достаточно долго следовать указаниям Экспериментатора (авторитета).

Причем же здесь данные?

Обратимся к евангелисту по цифровому маркетингу Google, Авинаш Кошик, который впервые ввел термин HYPPO в своей книге Web analytics: An Hour a Day.

HYPPO – означает мнение самого высокооплачиваемого человека в комнате (Highest Paid Person Opinion). Когда в комнате, где принимается решение, есть человек, который получает больше всех, то, скорее всего, его авторитет будет ключевым при формировании конечного решения.

Во многом такие решения могут противоречить тем, которые принимались на основании данных. Первые решения субъективны и, в конечном счете, преследуют личную выгоду, принося скрытый ущерб обществу. И как же быть? Ответ может лежать в плоскости деперсонификации принимаемых решений посредством анализа получаемых данных. Данные позволяют отказаться от эмоций и личной заинтересованности при анализе получаемых фактов.

Для этого процесс подготовки отчетности требует определенной реорганизации, как в прочем и самой организации.

7 шагов data-driven decision culture

В 2007 году, во время своего выступления в Google Conversion University, Авинаш Кошик выделил семь ключевых шагов, которые позволяют трансформировать культуру работы организации и перейти к дата-центрированной организации. И сейчас они не потеряли своей актуальности, поэтому я и привел их в этой книге как одну из основ построения новой формы культуры работы с данными.

Вот так называемые Cultural Hacks или Лайфхаки.

Шаг #1. Всегда переходите к Результатам – Go to the Outcomes

Основа коллаборации между людьми с использованием данных лежит, прежде всего, в понимании того, что важно для каждого из участников: от чего зависят их бонусы или выплаты, на что обращают внимание люди, которые принимают решения. Для этого нужно понимать, какими объектами оперирует компания, и это понимание перенести на уровень данных. Традиционная ошибка – начать собирать все данные компании, считать все возможные из них метрики и отправлять всем заинтересованным людям отчеты с этими показателями.

Шаг #2. Отчетность – это еще не Аналитика – Reporting is not Analysis

Большая часть отчетности, участвующая в подготовке, проверке или анализе, никак не связана с теми мотиваторами, от которых зависит завтрашний день каждого участника процесса, принимающего решение. В основном, ключевой ошибкой всегда и везде была простая демонстрация данных, в надежде, что решение с использованием этих данных найдет себя само.

На самом деле, в основе бизнеса лежат традиционные бизнес-вопросы.

Рассмотрим несколько примеров? Ведь это звучит это очень абстрактно.

Пусть у нас есть небольшая организация, где помимо прочих департаментов, есть целый отдел клиентского сервиса. Пусть вы являетесь топ-менеджером, уверен однажды это будет именно так.

Так вот, я утверждаю, что вы как руководитель будете

регулярно озабочены необходимостью постоянно задавать весьма конкретные и повторяющиеся вопросы о том, как обстоят дела с уровнем клиентского сервиса (успевает ли организация обслуживать своих клиентов вовремя и так далее).

Ответы на них будут лучше, чем просто отчетность, которая отвечает не на конкретный вопрос, а на открытый.

Все подобные вопросы можно выписать, структурировать и передать алгоритмам, чтобы они уже отвечали.

Шаг #3. Деперсонифицировать принимаемые решения – Depersonalise Decisions making

Переход к фокусировке на тех данных, которые действительно нужны организации, ведет к созданию новой формы культуры, где данным выделяют центральное место, а все решения – деперсонализированны, потому что важно не мнение людей в комнате, а данные на которых оно строится.

Нет смысла бороться с HYPPO, все решения должны быть деперсонифицированны, потому что они говорят не про мнения отдельно взятых людей, а про реальные тренды, бенчмаркинг, результаты работы клиентов или уровень их удовлетворенности. Будь-то электронная коммерция или реальное производство, данные покажут, что идет не в соответствии с ожиданиями, и это никак не связано с персональной оценкой.

А если HYPPO по стечению обстоятельств стал читатель этой книги, то для него важно помнить, что роль HYPPO – диверсифицировать мнение людей, допуская споры и несогласия. Своим присутствие HYPPO должен стимулировать принятие решений на основании данных.

Шаг #4. Проактивный инсайт (прогноз) важнее реактивной аналитики – Proactive insights rather than reactive

В тот момент, когда вы получили данные и начали заниматься подготовкой инсайта, данные уже устарели. Поэтому вместо того, чтобы выполнять и готовить отчетность, людям нужно выполнить анализ, про который никто не спрашивал ранее. Такой анализ необходим ввиду того, что данные быстро устаревают, и ряд ключевых аспектов может быть не покрыт во время процесса принятия решения.

Шаг #5. Расширить полномочия Аналитиков – Empower your

Analyst

Итак, для того чтобы Аналитик мог потратить свое рабочее время на анализ, о котором его никто не просил, у него должны быть достаточные полномочия, иначе, вместо подготовки регулярной отчетности, аналитик будет заниматься неструктурированным или слабоструктурированным анализом. Как ни странно, но data-driven организация вряд ли будет существовать в условиях регулярного процесса выпуска отчетности, на который тратится более восьмидесяти процентов времени работы команды. В одном из американских банков, где я однажды был на обмене опытом, была ситуация, когда люди выполняли регулярный процесс подготовки ежемесячной отчетности всего за 3 дня. Я спросил топ-менеджеров, а что люди делают остальное время, так как команда была достаточно большой. Они ответили – «Value Added активности», и все посмеялись. Признаюсь честно, до меня дошло не сразу. Под «делают Value Added активности» здесь подразумевалось, что аналитики использовали свое время, чтобы улучшить иные процессы организации по работе с данными и их продуктом – ежемесячной отчетностью.

Шаг #6. Треугольник ценности – Solve the Trinity

Внутри треугольника находятся метрики и инсайты, которые приводят к действию. На вершинах треугольника обозначены ключевые направления создания ценности с использованием данных:

• Поведение (Behaviour) – Необходимо думать широко при анализе поведения своих пользователей или клиентов. Это не просто данные, а поведение реальных людей.

• Результаты (Outcomes) – Научитесь связывать поведение клиентов с ключевыми показателями или критическими факторами успеха организации.

• Опыт (Experience) – Инсайты должны приходит через эксперименты, исследования, тестирование своих клиентов или поиск закономерности в их поведении. Этим необходимо постоянно заниматься.

Шаг #7. Создайте вокруг процесс – Got Process?

Data-driven организация – это не пункт назначения, а процесс или путь по которому идет организация, поэтому необходимо поддерживать его соответствующими артефактами и адекватными процессами. Этот процесс позволяет пользователям и сотрудникам применять тот или иной фреймворк работы с данными. Он не должен быть сложным и запутанным, а, скорее, должен отражать, кто и на каком конкретном шаге участвует в создании ценности с использованием данных.

Завершает Авинаш Кошик свой уникальный фреймворк одним из ключевых тезисов, без которого невозможно движение к data-driven организации, а именно: ответственным за данные, аналитику и поиск инсайтов в организации должно быть обособленное бизнес-подразделение (не IT).

В чем ценность data-driven организации

В 2011 году профессор MIT Эрик Брайнджолсфон провел любопытное исследование.[16] Он проанализировал данные 330 различных компаний за пятилетний цикл, в рамках которого выявил взаимосвязь между производительностью труда, выручкой и культурой организации, где было видно, как data-driven культура влияла на результативности той или иной компании.

Согласно исследованию, DD процесс повышал результативность труда и выручку компании на шесть процентов. По данным исследовательской компании Nucleous Research за 2014 год, было выявлено, что за каждый вложенный доллар в решения и процессы по аналитике и работе с данными, компания получала в среднем 13,01 долларов.

Data-informed организации

Продолжаем главы для продвинутых. Пытаясь разобрать дальнейший текст, я прошу, не сильно налегайте на алкоголь. Мне очень хочется, чтобы вы это прочитали.

Итак, существование так называемых дата-центрированных организаций имеет свое обоснование. Понятно, каким образом их строить и зачем. Но есть ли здесь какой-то подвох?

В 2010 году Адам Моссери, VP по продукту новостной ленты в Facebook, высказал мысль о том, как важно не допускать полной централизации организации в отношении данных. Основная идея его выступления сводилась заключалась в том, что данные дают возможность проанализировать текущую ситуацию и выбрать и наиболее оптимальный путь.

Но, если говорить о возможности создания уникального или лучшего продукта, то в дополнение к подходу, сформулированному Адамом Моссери, известный блогер и писатель в области Digital, Эндрю Чен, сформулировал тезис наличия «**локального максимума**[17]» в дата-центрированном процессе или продукте. Что это означает?

Локальный максимум представляет точку, которую можно легко выявить с помощью данных, и она помогает инкрементально (небольшими шагами) оптимизировать выбранный процесс или продукт. Но данная точка никак не связана с лучшей конфигурацией продукта или процесса, которая даст максимальный результат. Иными словами, при выявлении локального максимума всегда существует другая точка, которая является по совместительству экстремумом или наиболее лучшей конфигурацией продукта, но она отсутствует в наблюдении, так как данных для ее выявления обычно недостаточно.

Таким образом, путь развития организации как чисто дата-центрированной, перешел к новой модели работы с данными – data-informed.

Данная модель предполагает, что данные используются при принятии решений, но не являются ключевым фактором, так как поиск лучшего продукта является цепочкой экспериментов, которые заранее предсказать невозможно.

Каким образом сместить фокус с данных на другие аспекты, не потеряв важность работы с данными?

Ключевыми здесь всегда будут стратегия или видение того, что организация планирует делать. Так, в своем выступлении Адам Моссери, рассказал об оптимизации пользовательской функции по загрузке фотографии в Facebook. Его команда

провела ряд экспериментов по оптимизации процесса загрузки, руководствуясь при выборе того или иного интерфейса для пользователя только данными, начиная с кнопки и заканчивая изменениями во встроенных плагинах по поддержке браузера и навигатора для выбора файлов. Каждый из экспериментов оказался провальным, то есть не привел к увеличению конверсии активных пользователей сервисом загрузки фотографий.

В конечном счете, Моссери решил сменить тактику. Он оттолкнулся от данных, как стартовой точки анализа состояния воронки, и этапов, на которых пользователи по какой-то причине покидают Facebook. Затем он переработал подход, поставив во главу стола удобство пользователей и простоту.

Это дало определенные результаты, существенно увеличив конверсию пользователей. Конечное решение, выбранное его командой, не могло быть измерено только данными.

Data-informed или data-driven

При сравнении подходов ненамеренно вспоминается конфликт Стива Балмера (CEO Mircosoft 2000–2014) с Linux Foundation, которую он однажды назвал «раковой опухолью, приклеившейся к настоящей интеллектуальной собственности». В отличие от Microsoft, разработчик в Linux Foundation делает всего один патч для платформы за весь свой цикл работы на ней.

Данный конфликт очертил рамки нескольких типов организаций. По разные стороны виртуальных баррикад оказались разные подходы, в том числе и к управлению данными и инновациями.

Традиционный подход дата-центрированной организации опирался на правило Парето, которое гласит: двадцать процентов усилий приносят восемьдесят процентов результата. Высокопроизводительные силы сконцентрированы в дата-центрированных корпорациях, где есть нормативы, KPI, и где установка тех или иных требований к данным прямо влияет на получаемый результат или выполнение какого-либо норматива.

В дата-центрированных организациях основной упор в дизайне потребительских продуктов и сервисов строиться, прежде всего, на проверке гипотезы, где конечный потребитель (пользователь) голосует за наиболее приемлемый для него продукт, услугу или интерфейс.

Другой тип организации, наоборот, не имеет явных KPI или рычагов управления. Это так называемые организации открытого, платформенного типа. К ним можно отнести одно из ключевых утверждений, что дата-центрированные процессы не работают. С одной стороны, это пространство с неизвестными малоизученными переменными, где данные не могут однозначно повлиять на продукт, с другой, – этот тип организаций имеет одну отличительную черту, благодаря которой потребитель сам может стать создателем нового продукта или услуги. В таком случае сопутствующие аналитические сервисы, основанные на данных, позволяют потребителю самому создать для себя продукт который ему нравится.

В дальнейшем дата-центрированные организации могут использовать этот продукт для запуска на рынок. Так появилось много интересных продуктов, например, горные велосипеды, которые изначально придумали изобретатели в Калифорнии, переоборудуя специальные велосипеды со странным названием «балунеры» (или «кланкеры»).

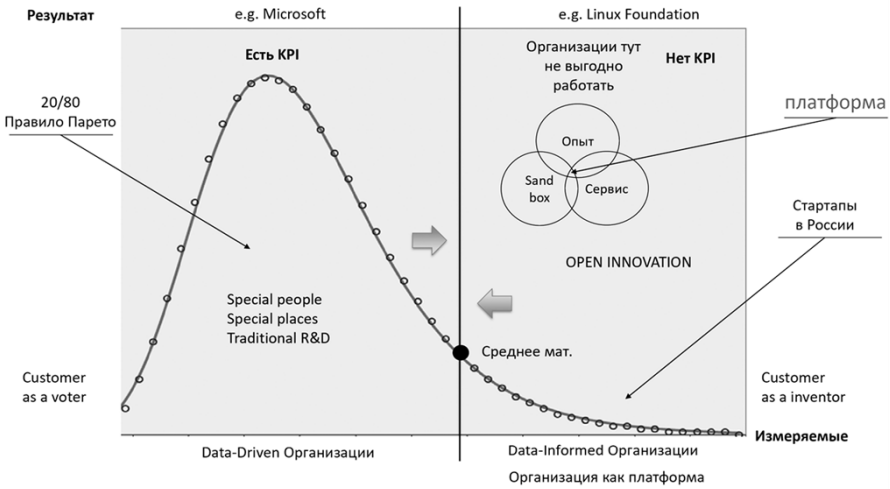
Важным фактом является то, что единороги, то есть

компании, капитализация которых измеряется в миллиардах долларов, появляются именно в организациях второго типа. Там, где нельзя ввести управление по показателям, а данные могут использоваться для сбора информации при принятии решения. Изучение long tail («длинного хвоста»), например в банкинге, является обязательным в надзорном регулировании. По основному замыслу принципов управления капиталом, разработанных Базельским комитетом, именно long tail может принести организации так называемый unexpected loss, то есть убытки, которые невозможно было предвидеть. Иными словами, «Черный лебедь».[18] И для них нужно рассчитывать определенный размер капитала, но организация это должна сделать сама, так как регулятор (например, Центральный банк) этого сделать не может. По аналогии с unexpected loss, возможен также unexpected profit, когда вместо убытка организация может получить сверхприбыль.

Это и есть те самые единороги, появление которых невозможно предсказать, опираясь только на данные.

Отличить один тип организации от другого, помимо анализа самой формы, внутренних процессов и других артефактов, можно так же оперируя только аналитикой.

Правило Парето перестает работать для процессов или показателей, значения которых попадают ниже среднематематического от потраченных усилий или ресурсов. Это означает, что если организация пытается ввести измерение процессов, которые не приносят существенный результат, или нельзя явно выделить процесс, который дает существенный результат, то такой тип организации становится data-informed, который исключает такой тип организации как data-driven (или data-centric).



Цикл развития организаций

Революция open-source и доступность технологий

Доступность технологий перешагнула барьер возможных применений, обогнав существующий спрос, а также приблизила так называемую точку сингулярности, за которой невозможно просчитать или спланировать возможный сценарий применения технологий.

Если рассмотреть эволюцию решений с использованием искусственного интеллекта, то в качестве интересных наблюдений, сделанных директором по маркетингу сервисов компаний «Яндекс», Андреем Сербрантом, можно привести в пример историю алгоритма AlphaGo[19]. В конце 2014 эксперты оценивали возможность искусственного интеллекта победить профессионала в го как маловероятный факт.

Год спустя, в декабре 2015 года, профессиональное сообщество повысило шансы на победу, но для обучения всем возможным стратегиям искусственному интеллекту еще требовались десятки лет.

Всего восемь месяцев спустя алгоритм AlphaGo, разработанный в лаборатории DeepMind,[20] смог обыграть профессионала игры в го, обладателя 18–ти мировых трофеев и высшего девятого дана, лидера мировых рейтингов, Ли Седоля.

Разработка алгоритма заняла всего несколько лет, при этом алгоритм не просчитывал все возможные комбинации, он мог видеть только на 50 ходов вперед. Просчет всех возможных комбинаций требовал несоизмеримое количество вычислительных ресурсов, поэтому разработчики решили пойти другим путем. Они создали трехслойную сложную нейронную сеть, которая имитировала человеческую интуицию. При этом алгоритм AlphaGo делал по-настоящему оригинальные ходы. Например, ход номер 37 в третьей партии против Ли Сидоля был действительно неожиданным для большинства профессионалов. Когда разработчики заглянули в логику алгоритма, они увидели, что AlphaGo выбрал этот ход, так как он был маловероятным ходом с точки зрения человека. Вероятность, с которой профессионал мог совершить этот ход, составляла 1:10000. Ли Сидоль проиграл эту партию, но в следующей игре он совершил аналогичный ход под номером 76, который был так же маловероятным, но, по его утверждению, он был единственным. Фактически, Ли Сидоль скорректировал свое понимание игры го,

в которую он играл с раннего детства, и применил новую отличную тактику, которую никогда ранее не применял.

Роль AlphaGo здесь совершенно не заменима при понимании основ игры го, потому что алгоритм смотрит на нее не так, как человек. Алгоритму не важны получаемые очки, потому что выиграть можно всего лишь с перевесом в одно очко, что и делает алгоритм. В результате получается новая, так называемая «мягкая» тактика, когда алгоритм стремится не к максимизации очков, а к устойчивому равновесию.

Появление таких сервисов изменяет саму суть игры, позволяет по-иному взглянуть на нее, применяя более зрелые подходы, которым учит нас алгоритм.

Сам алгоритм состоит из трех основных слоев:

• Стратегическая сеть – слой, который перебирает в памяти результаты всех сыгранных партий;

• Оценочная сеть – слой, который оценивает эффективность текущих позиций;

• Поиск по дереву – слой, который прогнозирует наиболее ценный ход руководствуясь эффективностью.

Если разобрать инфраструктуру, на которой был построен алгоритм AlphaGo, то это не какой-то сложный вычислительный суперкомпьютер. Его обучение проходило на пятидесяти графических процессорах в облаке Google Cloud. Если соотнести с рынком, то пятьдесят графических процессов эквивалентны небольшой майнинговой ферме по добыче криптовалюты, а использование облачных технологий делает весь процесс максимально мобильным.

Все оценки экспертов о невозможности существования такого алгоритма были разбиты. Это означает, что точка сингулярности технологий, о которой так много говорили, находится ближе, чем все думали[21]. Сегодня в разработке находится множество проектов, которые качественно иным образом упростят взаимодействие человека с окружающим миром.

Как ни парадоксально звучит, но хоть AlphaGo и работает на данных, совершая ход, он может учитывать и иные перспективы. Это означает, что если рассмотреть алгоритм как организацию, она одновременно демонстрирует черты как data-driven, так и data-informed. Возможно, это то будущее, которое будет наиболее эффективным в условиях постоянно меняющегося мира.

4-я промышленная революция, или Почему человек больше не нужен для поиска инсайтов

Говоря о возросшей роли данных в построении организаций нового типа, нельзя не отметить фундаментальный труд экономиста и основателя World Economic Forum Клауса Шваба, согласно которому мы переживаем четвертую промышленную революцию, основанную на данных.

Данные, алгоритмы распознавания и нейронные сети – все это позволило изменить традиционные процессы, вытеснить из них человека как необходимый элемент для обработки информации.

Отличным примером этого может быть сервис Stafory «Робот Вера» или Intervio от команды PryTek, который находит потенциальных кандидатов на выбранную позицию, обзванивает их, проводит их опрос и делает оценку соответствия потенциального кандидата предлагаемой позиции с использованием основных методик управления людьми, такими как Big Five. Происходит это благодаря сбору данных из баз резюме, таких как HeadHunter или TrudVsem. Так что, процесс поиска и отбора кандидатов на определенные позиции, уже сегодня может проходить без участия человека. Intervio – наоборот представляет собой сервис, где соискатель просто рассказывает свою историю, отвечая на вопросы, которые заранее записаны в виде видео интервью, а программа обрабатывает изображение, голос и получаемый текст и выдает оценки по психотипу, навыкам, используя сложный алгоритм нейролингвистического анализа. Это такой специальный алгоритм, который позволяет машине понять смысл слов. Например, «я хмурый иду по осеннему лесу» и «я иду по хмурому осеннему лесу» – два похожих предложения, но смысл у них разный. Машины уже способны уловить разницу в этом смысле.

С одной стороны, это серьезная трансформация процесса процесс подбора и резкое снижение его стоимости, с другой – чтобы пользоваться таким процессом, организации необходимо быть готовой внедрять такие сервисы в режиме Plug and Play, постоянно подключая эффективные цифровые сервисы и заменяя привычные процессы, требующие участия человека.

Датчики, телеметрия, бесконечные потоки данных, формирующие океан информации, создали новую цифровую экосистему. В ней с повышением интеграции данных в текущие

процессы меняется и роль человека. На смену традиционным профессиям индустриальной экономики приходит запрос на новые навыки в отношении управления и интеграции данных. Рынок и трансформация модели конкуренции открывают новые ниши для небольших игроков, которые формируют основное давление на современные большие компании. Чтобы быть эффективным, бизнесу придется акцентировать больше внимания в своем развитии на создание адекватной инфраструктуры сбора и обработки данных, а также решить ряд важных задач. Среди них ключевую роль играют методология и стандартизация протоколов передачи данных, информационная безопасность, аудит и управление качеством данных.

Потому что какими бы продвинутыми ни были алгоритмы, все они отступают при встрече с аномалиями в данных, причина которых может быть в некачественной информации. Поэтому проектирование, зачистка, контроль и арбитраж целостности – это одни из самых важнейших задач, которые придется решать в новой цифровой экономике.

Переход к новой парадигме работы с аналитикой, данными и информацией потребует от организации более высокого уровня зрелости, а это означает, что бизнес будет вынужден решить невыполнимую задачу по обучению специалистов и интеграции новейших технологий работы с данными в кратчайшие сроки, изменив при этом роль и ответственность участников цепочки создания информационного контента.

В этой книге я разберу основные приемы и модели, которые можно применять при выполнении этих задач, и которые помогут ответить на этот вызов. Мы с вами проанализируем: как строить команду, как выглядят новые профессии и какие методы управления могут применяться. Я расскажу, как можно разобрать кейсы, и покажу, как спроектировал новые сервисы, которые смогут заменить традиционные аналитические записки или отчетность.

Глава 2
Стратегия данных

С чего начинается стратегия данных?

Стратегию данных каждый из ключевых менеджеров компании сегодня понимает по-разному. А некоторые ее вообще до сих не понимают. Оно и понятно, много букв. Это как вишенка на торте инноваций и технологий, в котором еще надо уметь разбираться, чтобы просто банально насладиться тем вкусом, который есть. В том числе по-разному ее понимают и ключевые игроки рынка, производители программного обеспечения, разработчики и архитекторы данных. Нельзя просто взять, собрать всех вместе и наивно полагать, что получится договориться о чем-то одном.

Жизненный цикл данных

Данные – это что-то непонятное, неопределенное, как бесформенный прозрачный кислород. Вроде есть, вроде важен, но с чего начать?

Но во всех взглядах есть общее ядро, которое разделяется каждым из участников и является одним из ключевых факторов выбора и реализации стратегии – это понимание цикла работы с данными. Я выделил несколько моделей, иллюстрирующих наиболее полный жизненный путь данных внутри организации.

Например, модель Малькольма Чисхолма[22] выделяет семь активных фаз взаимодействия с данными:

1. Data Capture – создание или сбор значений данных, которые еще не существуют и никогда не существовали в компании.

a. *Data Acquisition* – покупка данных, предложенных внешними компаниями;

b. *Data Entry* – генерация данных ручным вводом, при помощи мобильных устройств или программного обеспечения;

c. *Signal Reception* – получение данных с помощью телеметрии (интернет-вещей).

2. Data Maintenance – передача данных в точки, где происходит синтез данных и их использование в форме, наиболее подходящей для этих целей. Она часто включает в себя такие задачи, как перемещение, интеграция, очистка, обогащение, изменение данных, а также процессы экстракции-преобразования-нагрузки;

3. Data Synthesis – создание ценности из данных через индуктивную логику, использование других данных в качестве входных данных.

4. Data Usage – применение данных как информации для задач, которые должно запускать и выполнять предприятие.

Использование данных имеет специальные задачи управления ими. Одна из них заключается в выяснении того, является ли законным использование данных в том виде, в котором хочет бизнес. Это называется «разрешенным использованием данных». Могут существовать регулирующие или контрактные ограничения на то, как фактически можно использовать данные, а часть роли управления данными заключается в обеспечении соблюдения этих ограничений.

5. Data Publication – отправка данных в место за пределами предприятия. Примером может служить брокеридж, который отправляет ежемесячные отчеты своим клиентам. После того, как данные были отправлены за пределы предприятия, де-факто невозможно их отозвать. Неверные значения данных не могут быть исправлены, поскольку они уже недоступны для предприятия. Управление данными может потребоваться, чтобы помочь решить, как будут обрабатываться неверные данные, которые были отправлены инвесторам.

6. Data Archival – копирование данных в среду, где они хранятся, до тех пор, пока не понадобятся снова для активного использования и удаления из всех активных производственных сред.

7. Data Purge – удаление каждой копии элемента данных с предприятия. В идеале это необходимо делать из архива, так как реализация задачи управления данными на этом этапе жизненного цикла данных определит, что очистка действительно была выполнена должным образом.

При работе с описанной моделью стоит отметить важные допущения:

• «Жизненный путь» – не совсем корректный термин, потому что данные сами себя не воспроизводят, более близкое значение – «история данных», но предлагается его не менять, из-за того, что текущего значения придерживается большинство участников рынка.

• Данные не обязательно должны проходить все семь фаз взаимодействия.

• Фазы взаимодействия не обязательно выстраиваются в конкретную последовательность. В реальности фазы могут проявляться в хаотичном порядке.

• Часть профессионального сообщества так же использует аббревиатуру ILM (Information Lifecyle Management). Разница[23] между двумя понятия состоит в следующем:

Управление Информационным Циклом (ILM)	Управление данными
• Стратегия информационного цикла • Средства аналитики • Управление данными на предприятии • Управление информационными активами • Управление контентом предприятия • Доставка контента • Архитектура и технологии	• Управление данными • Качество данных • Мастер данные • Метаданные • Архитектура данных • Приватность и безопасность • Консервация и архивация данных

Иными словами, по одной из версий управление данными является подмножеством цикла управления информацией, а сами подходы по управлению информацией уже являются подходами по управлению знаниями (Knowledge Management) в организации.

Но стратегия управления данными сама по себе является самостоятельным звеном в этой сложной цепочке. Поэтому, даже не рассматривая всю цепочку управления знаниями, можно с уверенностью сказать, что стратегия управления данными несет в себе самостоятельную ценность.

Утомил? А представьте, что в этом всем копается множество людей, которые в буквальном смысле спорят о дефинициях, правилах и отношениях.

Миссия компании и данные

Итак, при построении стратегии, вслед за определением ключевых точек работы с данными, обычно выбирается традиционный путь создания и разработки любой стратегии:

• Определение стратегической позиции – ответ на несколько ключевых позиций во внутреннем и внешнем окружении компании (с точки зрения регулятора, конкурентов, ресурсов и так далее), в том числе декомпозиция и интеграция миссии и ключевых факторов успешности;

• Определение стратегического выбора[24] – ответ на несколько ключевых вопросов: как именно организация будет конкурировать? В каком направлении? Как организация достигнет выбранного направления?

• Оценка и выбор стратегии – ответ на выборы по приемлемости предложенной стратегии.

Это основы любого стратегического планирования, которое мы не будем разбирать в этой книге, поэтому про него лучше почитать отдельно. Если собрать все основные подходы, которые в том числе известны мне, то получается следующая картинка:

Ключевые фреймворки при подготовке стратегии данных для организации

[1] Образована от сокращения шести английских слов: Political (политика), Economic (экономика), Social (общество), Technological (технология), Environmental (развитие) и Legal (законность). Данный анализ направлен на выявление политических, экономических, социальных, технологических и юридических или законодательных аспектов внешней среды, которые могут повлиять на стратегию компании.

[2] Методика для анализа отраслей и выработки стратегии бизнеса, разработанная Майклом Портером в Гарвардской школе бизнеса в 1979 году. Методикой выделяются пять сил, которые определяют уровень конкуренции и, следовательно, привлекательности ведения бизнеса в конкретной отрасли.

[3] Методика для анализа бизнеса, фокусирующаяся на доступных ресурсах в конкретной отрасли.

[4] Матрица Ансоффа представляет собой поле, образованное двумя осями – горизонтальной осью «товары компании» (подразделяются на существующие и новые) и вертикальной осью

«рынки компании», которые также подразделяются на существующие и новые.

Одно из ключевых свойств данных, которое необходимо учитывать при проектировании стратегической позиции компании – тот факт, что данные являются не только активом, который необходимо монетизировать, но и обязательством, за которым необходимо крайне внимательно следить во избежание штрафов, издержек или рисков, на которые компания должна аллоцировать соизмеримые резервы.

Перекладывая цикл данных на бизнес-приоритеты (иными словами, декомпозируя бизнес-модель на сильные факторы в текущей конфигурации), получаем следующую матрицу:

	Данные как Актив			Данные как Обязательство		
	Стратегическая позиция	Стратегический выбор	Стратегическая оценка	Стратегическая позиция	Стратегический выбор	Стратегическая оценка
Data capture	Стратегия данных			Стратегия данных		
Data Maintenance						
Data Synthesis						
Data Usage						
Data Publication						
Data Archival						
Data Purge						

Стратегия данных и жизненный цикл данных

Матрица позволяет разобрать на компоненты ключевой путь создания ценности из данных.

Таким образом, всегда есть два типа стратегии, которые будут развиваться:

• Стратегия защиты – сводится к минимизации риска владения данными. Она разворачивается вокруг ключевых активностей, таких как комплаенс, регулирование, выявление мошенничества с данными и других. Защитная стратегия так же ставит ключевой упор на стандартизации, управлении и оперативном выявлении рисков.

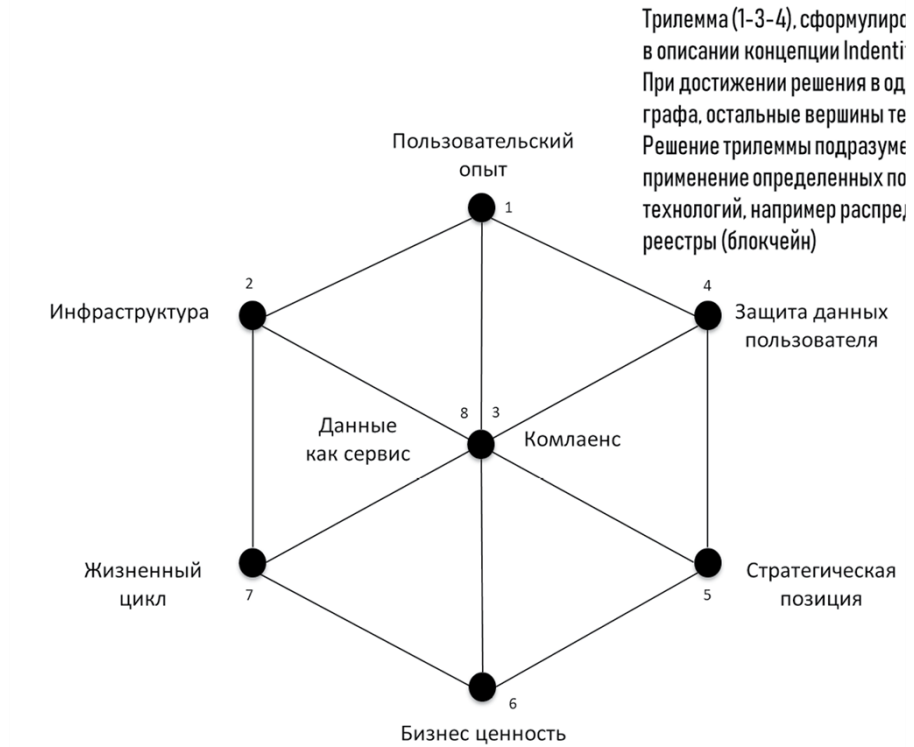
• Стратегия нападения – сводится к поддержке роста бизнеса (монетизации, росту конверсии и так далее). Ключевые активности обозначаются как новые знания о клиенте, поддержка решений и маркетинговые кампании.

Конечно, любой организации необходимо следовать обеим

стратегиям, но достижение баланса потребует от нее формулирование понимания аппетита к риску – единой позиции организации, так как эти две стратегии будут конкурировать за ресурсы самой организации. Огромное значение в этом будет иметь и размер самой организации, для таких моделей защитная модель всегда выглядит более предпочтительной, а для небольших компаний модель по поддержке роста бизнеса, наоборот, выглядит более преимущественной. Решения по выбору одной или другой всегда создают так называемый trade-off.

В качестве примера можно привести известную трилемму, сформулированную Яном Григгом (Ian Grigg).

Трилемма, сформулированная Ian Grigg в описании концепции Indentity-as-an-Edge. При достижении решения в одной из вершин, остальные вершины теряют ценность. Решение трилеммы подразумевает применение определенных подходов и технологий, например, распределенные реестры (блокчейн).



Трилемма Яна Григга и многомерная стратегия данных

Таким образом, стратегия данных имеет несколько измерений для анализа, каждое из которых необходимо учесть в соответствующей матрице по аналогии с тем, как это сделано для

вершин «жизненный цикл», «бизнес ценность», «стратегическая позиция».

Так, по данным **HBR**[25] выявлена, в том числе и зависимость от степени регулирования и выбираемой стратегии.

Ключевые стейкхолдеры

С точки зрения данных как актива, стратегия должна помогать использовать информацию в организации, поэтому основными стейкхолдерами стратегии в первую очередь должны быть зарабатывающие подразделения. С другой стороны, необходимо помнить, что успех во взаимоотношениях с клиентом лежит сегодня в том числе в возможности уметь рассказать о клиенте больше, чем клиент знает о себе сам.

Для данных как обязательства, помимо регулятора, есть еще бизнес-сообщество и клиенты, которым необходимо предоставлять актуальную информацию о соответствии требованиям законодательства. Например, новое европейское законодательство GDPR, вступившее в силу с 31 мая 2018, обязывает организации предоставлять конечным пользователям информацию и инструменты управления их данными.

Ключевые стейкхолдеры стратегии данных

У каждого этапа есть конкретный стейкхолдер, который может оказать наибольшее влияние на организацию. Их нужно учитывать в первую очередь.

Учет интересов большего числа стейкхолдеров позволяет

снизить издержки на коммуникацию и хранение данных в будущем, а также повысить шанс на их монетизацию.

Организациям, оперирующим на нескольких территориях, юрисдикциях или отраслях, необходимо учесть взаимное влияние на потенциальный размер рисков, которые создают специальные регуляции. Например, Общий Регламент по Защите Данных (или GDPR) применяет ряд следующих **принципов**[26][27]:

1) Законность, справедливость и прозрачность. Персональные данные должны обрабатываться законно, справедливо и прозрачно. Любую информацию о целях, методах и объемах обработки персональных данных следует излагать максимально доступно и просто.

2) Ограничение цели. Данные должны собираться и использоваться исключительно в тех целях, которые заявлены компанией (онлайн-сервисом).

3) Минимизация данных. Нельзя собирать личные данные в большем объеме, чем это необходимо для целей обработки.

4) Точность. Личные данные, которые являются неточными, должны быть удалены или исправлены (по требованию пользователя).

5) Ограничение хранения. Личные данные должны храниться в форме, которая позволяет идентифицировать субъекты данных на срок не более, чем это необходимо для целей обработки.

6) Целостность и конфиденциальность. При обработке данных пользователей компании обязаны обеспечить защиту персональных данных от несанкционированной или незаконной обработки.

Стоит отметить ряд важных аспектов, которые сегодня являются общими для значительного количества регуляторных юрисдикций (регуляций).

• Право на забвение, которое дает европейцам возможность удалять свои личные данные по запросу (во избежание их распространения или передачи третьим лицам).

• Право на переносимость данных (right to data portability) является новацией в правилах обработки данных ЕС, введенной GDPR. Данное право заключается в том, что компании обязаны бесплатно предоставлять электронную копию персональных данных другой компании по требованию самого субъекта персональных данных.

• GDPR устанавливает высокие требования в отношении формы получения согласия на обработку данных. Согласие человека на обработку его персональных данных должно быть выражено в форме утверждения или в форме четких активных действий пользователя. Согласие на обработку персональных

данных будет недействительно, если у пользователя не было выбора или возможности отозвать свое согласие без ущерба для самого себя. Если пользователь дал согласие на обработку своих персональных данных, контроллер должен иметь возможность продемонстрировать это.

GPDR не рекомендует использовать по умолчанию поля о согласии с уже поставленной галочкой или другие методы получения согласия по умолчанию. Согласие также не может быть выражено в виде молчания или бездействия пользователя. Информация о порядке отзыва согласия на обработку персональных данных должна быть размещена таким образом, чтобы пользователь мог легко ее найти.

Техническая инфраструктура

Стратегия выбора технологического стека, сопровождающего реализацию стратегии данных, во многом будет упираться несколько ключевых составляющих:

• Total cost of Ownership – совокупная стоимость владения технологией. Сюда попадают затраты как на сопровождение, так и на поддержку.

• Total cost of Change – совокупная стоимость изменений. Внешний мир меняется, поэтому в технологический ландшафт потребуется постоянно вносить изменения для того, чтобы соответствовать требованиям внешней среды.

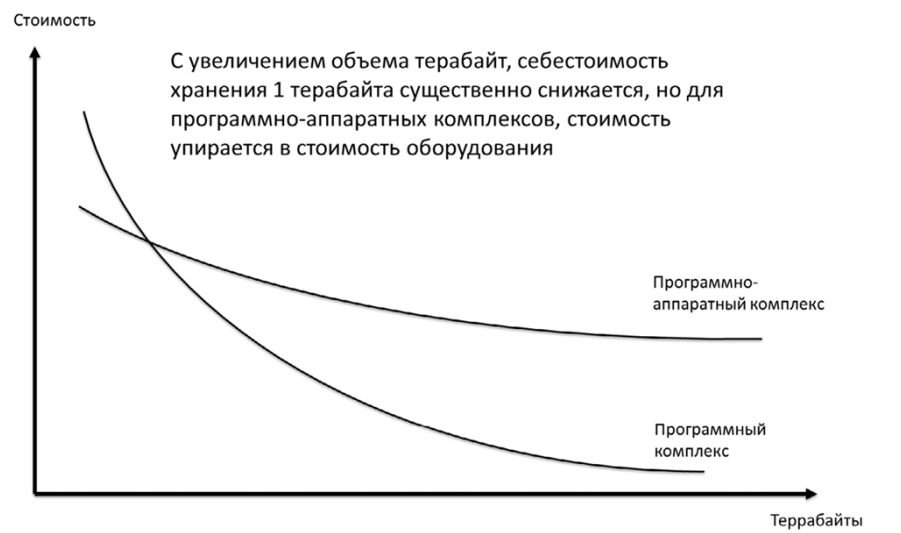
• Total cost per TB – совокупная стоимость технологии за один ТБ решения. При проектировании внутренней экосистемы одной из основных единиц цифровой экономики становится МБ, ГБ или ТБ хранения данных.

Выбор решения может подразумевать разделение на:

• Программные средства – софт, который позволяет обрабатывать или извлекать данные и проделывать с ними различные операции.

• Аппаратные средства – железяки, оборудование, без которого работа с большими и сложными вычислениями становится сложной и бесполезной.

• Программно-аппаратные средства – гибридные решения, которые включают в себя помимо оборудования также и софт.


Стоимость
С увеличением объема терабайт, себестоимость хранения 1 терабайта существенно снижается, но для программно-аппаратных комплексов, стоимость упирается в стоимость оборудования
Программно-аппаратный комплекс
Программный комплекс
Террабайты

Себестоимость владения технологией в зависимости от типа

средств

С точки зрения поведения экономики гибридные решения имеют ряд определенных ограничений при масштабировании стоимости такого решения и приобретении большего количества данных. В отличии от лицензий, ограничение вводят аппаратные средства, себестоимость производства которых зависит от конкретных драйверов затрат, что в свою очередь сложно переложить на драйвер объема данных. Поэтому с точки зрения снижения TCO более целесообразно выбирать так называемые «serverless-решения».

Зачем нужна стратегия данных?

Стратегия позволяет систематизировать усилия организации, прилагаемые для сбора и обработки данных, выделить ключевые события, которые происходят в данных, определить роли и участников и, наконец, поддержать реализацию миссии организации.

Lean цикл данных

Для простоты описания можно выделить несколько ключевых этапов, которые могут развиваться обособленно:

• Data Harvesting – эффективный сбор данных;

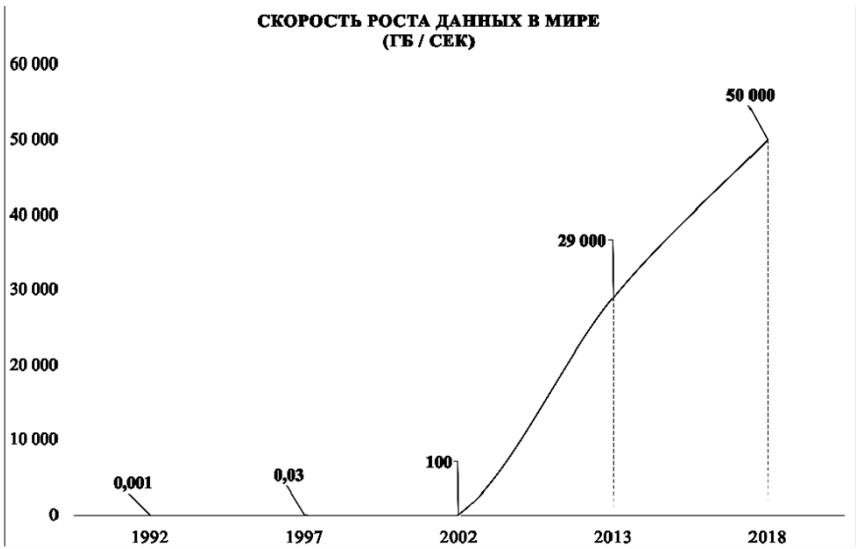
• Data Market – эффективный поиск и предоставление данных внутри организации;

• Data as a Service – аналитические сервисы с использованием данных.

Как влияет культура компании на успешность стратегии?

Согласно высказыванию Питера Дрюкера[28], «культура компании съедает ее стратегию на завтрак».

Если в компании ценность использования данных не поддерживается на каждом из уровней управления, то реализация стратегии данных находится под существенным риском.



Скорость роста общемирового объема данных

Большая часть данных, доступных сегодня для исследований и монетизации, была создана или собрана за последние несколько лет, и данные продолжают расти со скоростью 50 000 ГБ/сек[29]. Такие объемы дают определенное конкурентное преимущество при построении новой бизнес-стратегии, но лишь малая часть компаний успела трансформировать свою внутреннюю культуру для эффективной работы с данными на всех уровнях.

В свою очередь, технологии для обработки и хранения данных стали максимально доступными и экономичными. К 2050 году, когда население Земли достигнет 9,6 миллиардов человек, все объекты вокруг будут связаны в единую сеть обмена данными.

Быть успешной компанией, уметь рассчитывать показатели, планировать и ставить задачи стало невозможным без взаимодействия с данными, – с учетом того, что в скором времени большая части таких сервисов станет «коммодити»[30]. Чтобы

достичь этого, с одной стороны, организации необходимо выровнять единое понимание, что такое данные и какую ценность они создают для конкретной компании. С другой стороны, работа с данными требует скрупулезности и аккуратности. При развитии навыков и компетенций работы с данными, например при машинном обучении, происходит обособление от такой науки как статистика. Возникают постоянные барьеры коммуникаций, сводящие к минимуму возможность успешной кооперации.

Такие барьеры приводят к одной из важнейших проблем в управлении данными – департаменты внутри корпораций работают в формате «Silos»[31] – с изолированными хранилищами данных, которые возникают естественным образом в крупных организациях. По сути, речь идет о «подстольном» BI-хранилище, которое стоит практически у каждого отдельного департамента, и, как показала моя практика, это явление весьма частое. Такие Silos делают невозможным достижение и создание «Единого хранилища данных».

Silos возникают, когда департаменты конкурируют друг с другом. Важно понимать, что основу такой конкуренции создает внутренняя культура организации, поэтому стимулирование внутренней конкуренции вредит стратегии данных. Можно даже утверждать, что вероятность совместить такие организации, где поддерживается и стимулируется конкуренция со стратегией данных, крайне низкая.

Silos как явление существуют не только внутри организации. Если рассмотреть несколько отраслей, например, производство и банкинг, то здесь данные изолированы и хранятся только внутри производственного контура. Банк с наименьшей вероятностью сможет получить доступ к данным производства, хотя как раз получение данных дает возможность разработки и создания «цифрового двойника» производства и моделирования новых финансовых продуктов с использованием данных, таких как гарантии или производственный овердрафт, без необходимости сбора бумаг или отчетности. Именно поэтому фактор культуры и устранения барьеров коммуникации является одним из ключевым при построении дата-центрированной бизнес-модели.

ПЛАН ПО УСТОЙЧИВОМУ УРОВНЮ КАЧЕСТВА ДАННЫХ	ПОЛНОТА	АКТУАЛЬНОСТЬ	ДОСТОВЕРНОСТЬ	ДОСТУПНОСТЬ
КЛИЕНТСКИЙ ДОМЕН	99,99%	95%	95%	95%
Вес в общем показателе	40%	30%	15%	15%
ПРОДУКТОВЫЙ ДОМЕН	99,99%	99%	99%	99%
Вес в общем показателе	40%	40%	5%	15%
ДОМЕН «СТРУКТУРА»	99,99%	99%	99%	95%
Вес в общем показателе				
DATA LINEAGE [загрузка, трансформация] Допустимая деградация	- 1%	- 1%	- 1%	- 1%

Пример моей стратегии по управлению качеством данных на основе выделенных доменов (блоков данных)

Ряд экспертов[32] предлагает несколько решений по гармонизации и трансформации культуры организации:

• **Открытость** – сотрудников стимулируют делиться данными, высказывать идеи и поощряют за помощь в исследованиях данных других департаментов и за их использование.

• **Top-down** менеджмент напрямую координирует и на своем примере показывает важность совместной работы с данными.

• **Холократичность** – сведение всех ключевых заинтересованных лиц в круг влияния по примеру компании Zappos; формирование «плоских» или одноранговых команд для работы над аналитическими сервисами с использованием данных.

• **Стройте сервисы** – переход на сервисную модель работы с данными, позволяющий стандартизировать и выравнивать понимание того, как должен выглядеть тот или иной сервис.

• **Фокусируйтесь на драйверах**[33][34] – определение драйверов, которые приводят к появлению Data Silos, такие как:

◊ **Множество и различие платформенных решений и компетенций** – когда внутри одной организации существуют одновременно много различных платформ по работе с данными.

◊ **Политические** – борьба за сферы влияния приводит к тому, что информация используется как основной инструмент для разделения влияния.

◊ **Неравномерный рост** – быстрый рост компании или неорганические приобретения различных бизнес групп приводят к тому, что возникают отличные интерпретации того, как использовать данные.

◊ **Сфокусированность на вендоре** – каждый из вендоров имеет внутри своего решения уникальную модель данных. Многие из них строят изолированные экосистемные решения, которые не умеют находить общий язык с решениями других вендоров. Сегодня стандартизированы только интерфейсы без интерпретации.

По версии Digital Impact[35] предлагается, наоборот, рассмотреть ряд нестандартных приемов по трансформации культуры организации:

• Предложить сотрудникам **делать скетчи с историями** про данные. Сотрудники изучают данные и пробуют рассказать историю, для этого организуются регулярные питчи внутри компании в специально отведенное время (так называемые DemoDay).

• **Построить скульптуру данных**, которая будет представлять те или иные данные. Необходимо подумать и сконструировать решение, которое в том числе будет привлекать внимание других сотрудников и поможет впоследствии рассказать историю #datasculpture.

• **Начать формулировать аргументы** с использованием данных во время дискуссии или обсуждения.

Кто владелец стратегии данных?

Анализируя структуру навыков и требований к современному **Data Scientist**[36][37] (которая, кстати, уже тоже устарела, так как на смену работе с Hadoop пришел Spark для работы с NoSQL БД), можно выделить ряд ключевых ожиданий.

Современный исследователь данных по версии MarketingDistillery

Помимо навыков из области математики или статистики, специалист в обязательном порядке должен обладать навыками, позволяющими ему уметь настраивать среду, загружать и обрабатывать данные и подготавливать датасет к исследованиям.

Кроме специалиста по исследованиям необходимы специалисты по контролю и качеству данных, бизнес-аналитики, архитекторы данных, разработчики информационных потоков и сервисов и так далее.

При всем обещающем многообразии компетенций и навыков встает открытый вопрос: к кому из бизнес-лидеров должна относится стратегия данных. Один из традиционных взглядов, преобладающий в большинстве компаний, выглядит следующим образом:

• **Финансовый директор** отвечает за стратегию данных и имеет специальное подразделение, которое выступает заказчиком и контролирует качество данных для всех остальных

подразделений.

• **IT-директор** отвечает за реализацию, наполнение и сбор данных в соответствии с выставленными требованиями (SLA, OLA и так далее).

В такой конфигурации возникает несколько коллизий при работе с данными:

• **Финансовое подразделение** оперирует размерностями данных, которые в первую очередь будут покрывать потребности подразделения, входящие в зону его ответственности перед внешними инвесторами (МСФО отчетность, Investor Relations и другие). В этом смысле многомерная и сложная бизнес-сущность организации представляется в виде плоского отчета, во много отвечающего ограниченному количеству аналитических задач.

• **IT-подразделение** не берет на себя ответственность за качество данных в источниках. Помимо этого, гармонизация источников данных также требует приложение сверхусилий[38].

Решать такие коллизии призвана модель офиса CDO (Chief data officer) в прямом подчинении CEO, в котором появляются ряд новых профессий и ролей – например, **data engineer**[39][40] или data architect. Они вместе с CDO проектируют и внедряют ряд ключевых артефактов, на которых будет строиться стратегия управления данными. Это могут быть:

Восприятие организации с помощью данных

ИНЖЕНЕР ДАННЫХ
(DATA ENGINEER)
ОБРАБОТКА СЫРЫХ ДАННЫХ
СОЗДАНИЕ ИНФРАСТРУКТУРЫ ДЛЯ КОНСОЛИДАЦИИ И ОБОГАЩЕНИЯ ДАННЫХ
ОБРАБОТКА И УПРАВЛЕНИЕ СИСТЕМА И ПОТОКАМИ ДАННЫХ
ПОДГОТОВКА ДАННЫХ ДЛЯ АНАЛИЗА
ИССЛЕДОВАТЕЛЬ ДАННЫХ
(DATA SCIENTIST)
ПОИСК ПОЛЕЗНОГО КОНТЕНТА В ДАННЫХ НА ОСНОВЕ МЕТОДОВ
ПОДГОТОВКА РЕЗУЛЬТАТОВ ДЛЯ БИЗНЕСА
ИНТЕРПРЕТИРОВАНИЕ И ДЕКОДИРОВАНИЕ ДАННЫХ
ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ РАБОТЫ

Отличие инженера данных от исследователя данных

• Единая бизнес модель и единая модель данных.
• Аппетит к риску на основании.
• Data Quality и так далее.

В своей основе data engineers имеют ряд отличительных особенностей от data scientists, если поставить их в один ряд, то можно сказать, что data engineers больше занимаются самими данными, нежели поиском инсайтов из них. Их задача – следить, проектировать и организовывать бесконечные потоки данных, структурируя и валидируя их для конечного пользователя.

Self-service BI

Отдельно стоит рассмотреть экосистему Microsoft, организованную для двухсот тысяч сотрудников корпорации, и предоставляющую все необходимое для работы с данными. Вызовы, на которые отвечает эта экосистема, сопоставимы с задачами по трансформации культуры, стоящими перед крупнейшими корпорациями.

Команда Microsoft выделила пять видов особенностей в реализации стратегии данных:

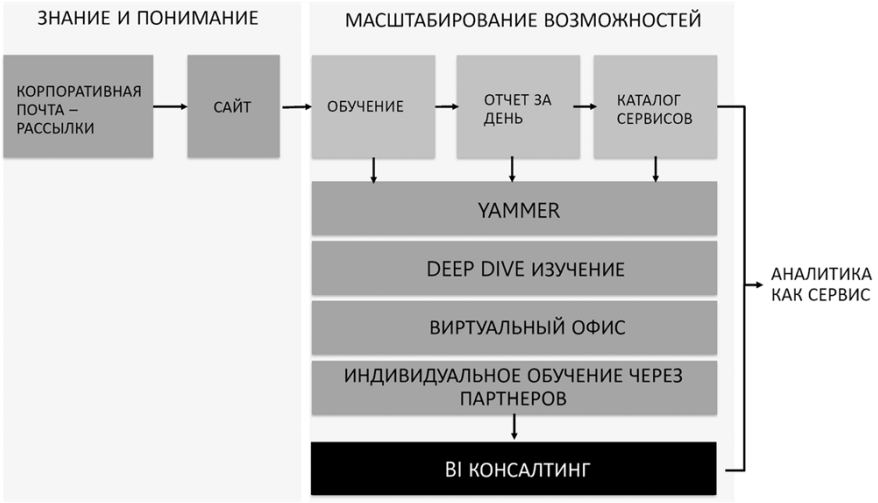
• Заменить стратегии оценки эффективности внедрения BI средств на стратегии возможности взамен того, чтобы пытаться оценить ROI от проектов, связанных с данными, организация должна перейти к пути оценки возможностей применения данных.

• Перейти от управления изменения (Change Management) к модели потребления данных. Сервисы на данных – это продукт, у которого есть свой потребитель. Технологическая организация должна полностью сфокусироваться на потреблении технологических продуктов.

• Сфокусироваться вокруг кривой использования BI-инструментов и ранних последователях (Early Adopters), так как они являются самыми важными бизнес-пользователями, которые будут потреблять тот или иной сервис.

• Структурировать инструменты поддержки для каждой группы пользователей с точки зрения канала коммуникации, поддержки продукта, общего видения развития сервиса и так далее.

• Сформировать экосистему поддержки инноваций и работы с данными с вовлечением социальных сетей, каналов коммуникаций, партнеров и поставщиков данных, создавая возможность быстрого масштабирования.



Инфраструктура Microsoft по поддержке развития BI-сообщества

Итак, комплексность взгляда Microsoft на управление культурой данных в больших корпорациях показывает, что помимо трансформации понимания роли данных (перед от ROI и других показателей к оценке возможностей), от организации требуется также глубокая и детальная проработка инструментов поддержки жизненного цикла данных, сегментации потенциальных потребителей и выделение ресурсов на продвижение и поддержку каналов.

В этом смысле управление и развитие таких инициатив сопоставимо с развитием и созданием нового бизнеса, где данные и сервисы на них являются продуктом, а пользователи становятся полноценными потребителями.

ЦИКЛ BI ВНУТРИ ОРГАНИЗАЦИИ
по версии MICROSOFT
Изучаем основы – критерии успеха, как применять и коммуницировать с аудиторией
Важно быстро наращивать темп развития – строить по 1 дэшборду в день
Высокий запрос на поддержку и быстрый доступ к данным
ИССЛЕДУЕМ МЕТОДАМИ ПРОБ И ОШИБОК
Желание улучшить и оптимизировать персональный опыт BI
Поиск лучших практик, применение утвержденных методоий MICROSOFT, выход из модельности "Just Do it" в модальность "Do It Right"
Демонстрация кейсов, формирование внутренней продажи аналитики как сервис для ТОП-Менеджмента с использованием КПЭ (экномия или новый доход)
РАЗВИТИЕ РОЛИ BI ВНУТРИ ОРГАНИЗАЦИИ
Декомпозиция личных целей топ-менеджмента с использованием BI инструментов – их личное вовлечение и инвестиции
Ощущение успеха становится частью внутреннего сообщества инноваторов
СОЗДАЕМ ИННОВАЦИИ

Путь формирования культуры работы с данными, по версии компании Microsoft

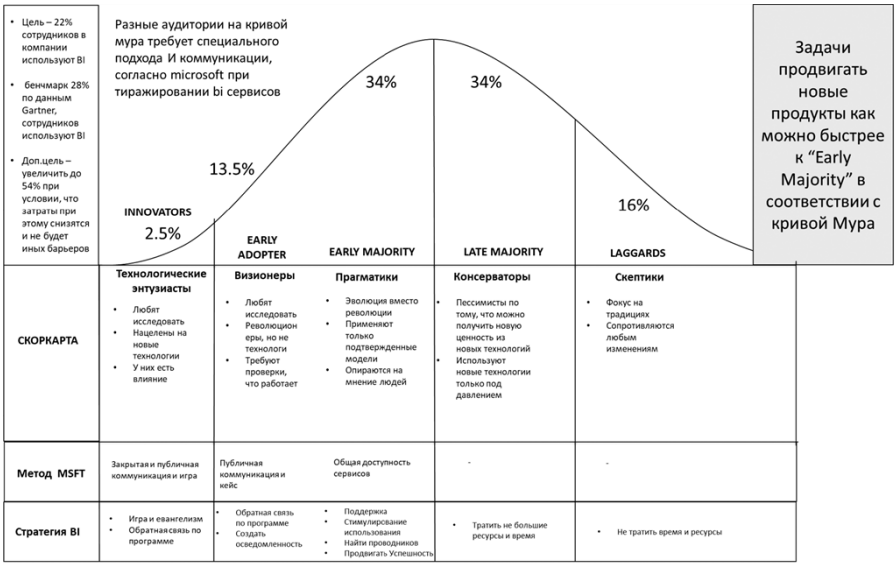
Известная кривая Мура[41] определяет группы пользователей по взаимодействию с технологией. Ею пользуются большинство компаний в Силиконовой Долине, потому что она содержит ключевую подсказку.

Все пользователи делятся на две группы по пятьдесят процентов. Первая группа имеет явно выделенные внутри три подгруппы:

• **Инноваторы** – они составляют всего два с половиной процента от общего количества возможных конечных пользователей аналитического продукта. Они ищут новые знания, хотят попробовать новые технологии, им важно влиять и менять новые, зарождающиеся технологии.

• **Ранние последователи** – их уже больше тринадцати с половиной процентов от общего количества возможных конечных пользователей. Они являются визионерами, поэтому не будут обращать внимания на «шероховатости» аналитического продукта. Тут возникает знаменитое правило шестнадцати процентов о том, что первая часть аудитории быстро соглашается на изменения, поэтому им легче «продать» или объяснить новые технологии и подходы. Культура работы с данными в этом – не исключение, как показал пример Microsoft. После шестнадцати процентов возникает так называемый «Разрыв», а именно, бетонная стена, в которую врезаются все инноваторы, так как следующая категория ребят уже хочет работающий аналитический сервис, а значит, они не готовы больше проглатывать все шероховатости.

• **Раннее большинство** – их уже больше, целых 34 процента. Это настоящие прагматики, которых убедят использовать продукт только их друзья или доверенные лица, но если они перейдут на него, то будут продвигать успешность этого продукта.



Кривая Мура об адаптивности инноваций, на примере программы Microsoft об активации культуры работы с данными

По версии Мура, каждая группа требует определенного послания или коммуникации. И все эти послания разные, как можно понять. Microsoft, понимая это, разработал уникальную систему коммуникаций, где явно отделил одних пользователей от других и структурировал послание, которое он несет для каждой из групп. Поэтому, если вдруг внутри вашей компании вы внезапно начнете нести проповедь про культуру данных, нужно помнить, что услышать ее могут далеко не все, а только два с половиной или тринадцать процентов, если уже будет что показать.

Как измерить успешность стратегии данных?

Команда некоммерческой лаборатории Digital Impact[42] определила следующие критерии успешности реализации стратегии данных:

• Сотрудники распознают, что такое данные, когда их видят, и предлагают креативные решения по их использованию.

• Сотрудники поддерживают и предоставляют доступ к совместному использованию данных.

• Менеджмент организации инвестирует время и средства в развитие инструментов по сбору и анализу.

С другой стороны, измерение стратегии данных потребует формулирования ключевых факторов успеха, необходимых для реализации стратегии (Key Success Factors). Их достижение будет означать успех в реализации стратегии данных. Например, одним из таких факторов может быть поддержание качества данных на определенном уровне.

Качество данных можно измерить разными способами и разными показателями, такими как:

• **Полнота** – количество данных в источнике и хранилище (или в отчете или в любом другом месте) совпадает. Нет материальных искажений в полноте описания совершившихся транзакций.

• **Актуальность** – все описанные события актуальны, то есть произошли в релевантном временном периоде.

• **Достоверность** – каждое из событий существует в реальном мире и может быть подтверждено соответствующим документом, сотрудником или независимым участником.

• **Доступность** – ко всем необходимым данным имеют доступ соответствующие сотрудники, все важные атрибуты и транзакции, формируемые в бизнесе, могут быть получены.

Сколько стоит реализовать стратегию данных?

Реализация всегда затрагивает несколько ключевых измерений:

• Технологии
• Людей
• Процессы

В каждом из измерений необходимо сформулировать те самые критерии успешности, к которым будет стремиться организация.

Технологии

Выбор подхода к созданию внутренней экосистемы будет влиять на себестоимость хранения одного терабайта. Ценообразование Enterprise Grade решения (для корпоративных систем) стоили кратно дороже, чем стоимость решений на open-source.

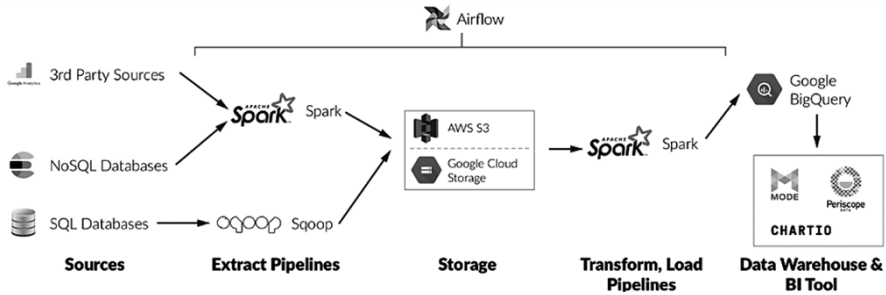
Например, в 2012 году Fusion Alliance дал оценку в среднюю сумму по рынку с учетом дисконта в шестьдесят процентов (скидка от основной цены предоставляемой вендором), которая составила 26 тысяч долларов за один **ТБ**[43][44], против четырехсот долларов затрат на ТБ для решений на открытом коде. В эти затраты входили лицензии, установка и разработка, закупка и настройка необходимого оборудования.

Минимальная конфигурация шла в составе шестнадцати ТБ, таким образом, планирование происходило блоками. Позднее, в 2015 году, себестоимость начала снижаться с появлением программных комплексов (таких как HP Vertica) до пять тысяч долларов, без учета затрат на оборудование.

Сегодня создание и управление облачным хранилищем данных, например, с использованием сервисов Google или AmazonWebServices, будет обходиться существенно меньше – от десяти до сорока долларов за все.

Проект Apache сегодня насчитывает порядка 38 различных решений с открытым кодом по Big Data, ряд из них конкурируют, но большая часть решают индивидуальные задачи.

Так **Нейт Кнапп**[45],инженер компании Thumbstack, предлагает рассмотреть следующую экономичную инфраструктуру из компонент с открытым кодом:



Архитектура Big Data решения от Нейта Кнаппа с использованием бесплатных решений обработки данных

• Spark – фреймворк с открытым кодом для реализации распределенной обработки и загрузки неструктурированных данных.

• Scoop – фреймворк с открытым кодом, предназначенный для обработки и передачи данных между структурированными и неструктурированными источниками данных.

• Airflow – фреймворк-планировщик, который управляет сессиями задач для фреймворков загрузки данных. Разработан компанией AirB'n'B.

• AWS / Google Cloud Storage – облачный сервис по хранению данных.

• Mode, Periscope, Chartio – платформы и фреймворки по визуализации данных и подготовке отчетов.

Большая часть из них относительно бесплатны в использовании.

Люди

Ключевые затраты, связанные с персоналом, разделяются на ряд ключевых факторов:

• **Обучение специалистов** новым технологиям. Сегодня объем новых доступных технологий и распределение навыков в профессиональной среде слабо коррелирует, поэтому необходимо вкладываться в постоянное развитие и обучение специалистов.

• **Поиск и найм** талантливых кадров. Из-за высокой конкуренции за ресурс, в среднем по рынку срок работы на одном месте высококвалифицированного специалиста – менее трех лет, поэтому необходимо постоянно работать в направлении привлечения и удержания ключевых сотрудников, создающих ценность из данных.

• **Новые профессии и структуры.** Создание офиса CDO

может столкнуться внутри организации с радикальными трансформациями. Не всегда существующие на рынке практики позволяют применить эти модели внутри организации. Вполне возможно, что потребуется создание новых профессий.

Выбор стратегии в части персонала будет зависеть от характера CDO.

Например, в части цифровых технологий, в одном из исследований специалисты компании **PwC**[46],сформулировали, что существует пять архетипов CDO (Cheif Digital Officer):

• **Прогрессивный мыслитель (Progressive Thinker)** – миссия этого исполнительного лица состоит в том, чтобы думать, как бизнес может быть преобразован посредством цифровизации и вдохновлять компанию на полноценную цифровую стратегию и операционную модель. Желательно работать непосредственно с генеральным директором. Прогрессивный мыслитель, как правило, сосредотачивается на разработке цифровой стратегии и инноваций в масштабах всей компании, одновременно преследуя новаторские идеи и методы как в мире в целом, так и в отрасли. Компании в более традиционных отраслях промышленности, таких как химикаты, нефть и газ, а также горнодобывающая промышленность, которые уже имеют стабильный и сильный набор дифференцирующих возможностей, но до сих пор не извлекли полной выгоды из оцифровки, должны рассмотреть вопрос о найме прогрессивного мыслителя. Это CDO, который может принести вдохновение и опыт, необходимые для дальнейшей оцифровки текущей стратегии компании.

• **Креативный дизраптор (Creative Disrupter)** – в отличие от прогрессивного мыслителя, «творческий разрушитель» предлагает более практичный подход к непрерывной разработке новых цифровых технологий, бизнес-моделей и решений. «Творческий разрушитель» может быть особенно ценным в компаниях, сталкивающихся с серьезными и драматическими изменениями в результате оцифровки – в отраслях, ориентированных на потребителя, таких как публикации и розничная торговля. Ему или ей придется работать непосредственно с генеральным директором по бизнес-ориентированному подходу к созданию конкурентной дифференциации, которая приведет к росту доходов и повышению прибыльности, часто путем включения идей и технологий извне их традиционных отраслевых структур и конвенций.

• **Адвокат клиента (Customer Advocate)** – так называют руководителей, которые обычно отчитываются перед CMO и руководителем продаж или могут даже заменить их, в основном они ориентированы на рынок и удовлетворение потребностей

клиентов. В результате адвокаты клиента лучше всего подходят для компаний в отраслях, ориентированных на интересы клиентов: розничная торговля, банковское дело и путешествия, особенно если цифровое мышление еще не проникло в повседневную жизнь их продавцов и специалистов по маркетингу. Адвокат клиента фокусируется на разработке удобного, привлекательного и бесшовного опыта работы с клиентами, используя дизайнерское мышление по всем каналам – цифровым и физическим. Таким образом, он или она несет ответственность не только за онлайн-функцию продаж, но и за последующее обслуживание и послепродажное обслуживание. Компании с портфелями продуктов, уже дифференцированные путем оцифровки, могут ограничить роль CDO-защитой для клиентов и опытным трансформатором.

• **Инновационный технолоджист (Innovative technologist)** – этот CDO, как высокоинновационный и ориентированный на бизнес IT-директор или главный технический директор (CTO), продвигает использование новых цифровых технологий для преобразования цепочки создания стоимости всей компании, обеспечивая технологическую основу для новых цифровых бизнес-моделей с помощью таких технологий, как Интернет (IoT), мобильности, социальных сетей и аналитики, а также повышения внутренней эффективности и поиска путей сокращения издержек. Цель инновационного технолога заключается не в том, чтобы нарушить работу компании, внедряя способы ведения бизнеса из других отраслей, а в том, чтобы скорее работать на границах отрасли, используя цифровые технологии для получения конкурентной дифференциации за счет скорости, эффективности и развития новых бизнес-моделей, основанных на объединении цифровых услуг с физическими продуктами. Например, компании в обрабатывающей промышленности должны рассмотреть возможность обращения к этим руководителям для дальнейшей оптимизации своих цепочек поставок и внедрения цифровых технологий на заводы для таких ключевых этапов производства, как проектирование и прототипирование. В отличие от Industrial Internet или Industry 4.0, оцифровка производства окажет значительное влияние на эти компании, и инновационный технолог будет играть важную роль в их способности извлечь выгоду из этой трансформации.

• **Универсалист (Universalist)** – миссия этого типа CDO, как правило, заключается в управлении всеми аспектами и рычагами создания ценности, полной цифровой трансформации. Самый дальновидный из пяти архетипов, универсалист, может добиться успеха только благодаря сильному мандату от генерального

директора и полной поддержке исполнительной власти. Отчитываясь непосредственно полностью поддерживающему его генеральному директору, универсалист может контролировать спектр возможных цифровых задач: руководство разработкой цифровой стратегии в рамках общей корпоративной стратегии; разработка новых бизнес-моделей; надзор за цифровым маркетингом и опытом работы с клиентами; внедрение цифровых технологий; повышение операционной эффективности; он может также оцифровывать способы, с помощью которых сотрудники выполняют свою работу. Универсалист должен отвечать за процесс управления организационными и культурными изменениями. Этот архетип особенно хорошо подходит для компаний в любой отрасли, которые оказываются за поворотом в своих усилиях по адаптации к цифровому миру и поэтому нуждаются в исполнительной власти, которая может провести быструю и всеобъемлющую трансформацию.

В зависимости от того, какой из указанных архетипов подходит к организации, будет сильно меняться сама управленческая парадигма – модель управления человеческим капиталом. Каждый из указанных архетипов, со стороны PwC, определяет различные модели управления:

• **Подчинение** – прямое под CEO, или вхождение в состав CMO, CIO или CTO.

• **Зоны ответственности и компетенций** – множество различных задач для каждого из архетипов, начиная от разработки цифровой стратегии, заканчивая поддержкой клиента в CRM-системе.

• **Бизнес-результат и KPI** – CDO может иметь как явный бизнес-результат, так и неявный, и быть лишь центром затрат с соответствующими показателями оценки эффективности деятельности.

• **Платформы и компетенции** – в зависимости от модели будет также изменяться технологический ландшафт, например, необходимость включения CRM или IoT.

Процессы

Получение быстрого результата потребует от организации эффективного пост-пространства для креативной работы сотрудников.

Так, консультанты компании McKinsey предложили использовать Agile для формирования совместных эффективных небольших **Data Teams**[47].

Agile команда Data Lab по версии McKinsey

При этом так же упрощаются существенно сами этапы получения данных и инсайтов:

• Харвестинг данных (или сбор данных)

• Использование гипотез при исследовании

• Создание аналитических сервисов (продуктов на основании данных)

• Модель управления данными (Governance)

• Презентация полученных кейсов.

Первое, о чем стоим договориться команде, – как выглядит **Definition of Done**[48][49] по Аналитическому продукту или продукту с использованием данных.

Для организационных структур, которые требуют конкретного описания процессов, всегда доступны стандартные swim lane диаграммы, разработанные командами ведущих компаний.

Например, команда **Microsoft**[50] представила исчерпывающую методологию построения процесса изучения данных и получения исследований, опираясь на жизненный цикл данных и стандартизированную ролевую модель:

• Архитектор решений (solution architect)

• Руководитель проекта (project manager)

• Исследовать данных (data scientist)

• Руководитель проекта (project lead)

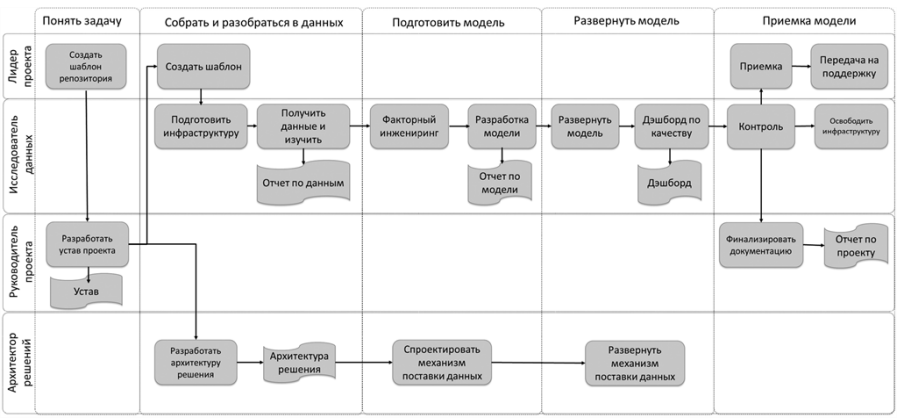
С другой стороны, для извлечения максимальной ценности и

получения адаптивного к моменту процесса управления данными, появилась методология DEVOPS, которая была сформулирована Энди Палмером (**Andy Palmer**[51]), CEO и сооснователем компании TAMR (ранее – основатель компании Vertica).

По его мнению, на ее распространение повлияло несколько ключевых факторов:

• Демократизация аналитики – сегодня все больше людей по всему миру работают с аналитикой.

• Создание специальных баз данных (Vertica, VoltDB, StreamBase, BigTable) под задачи – реляционные базы данных устарели, и сегодня одно решение не подходит для любых задач.



Процесс поиска инсайта по версии Microsoft

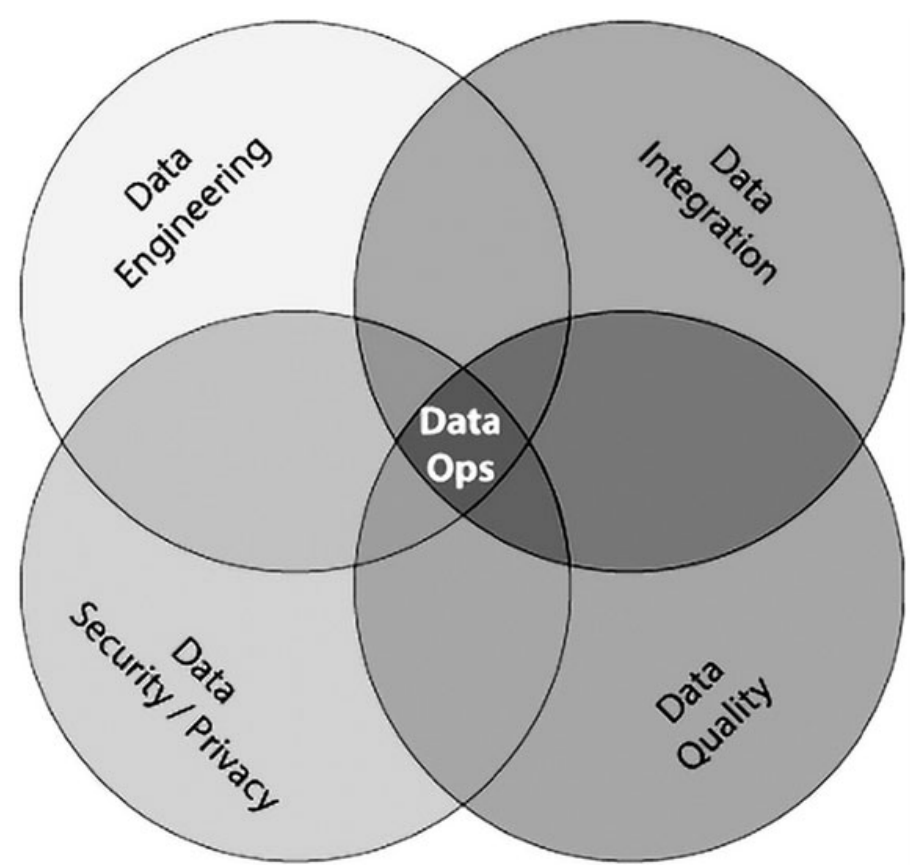
С одной стороны, решения перестали быть универсальными, с другой – решения должны иметь стандартные интерфейсы (API) для интеграции различных решений. Вместе эти тенденции создают «давление с обоих концов технологического стека».

В верхней части стека – все больше пользователей хотят получить доступ к большему количеству данных в большем количестве комбинаций. А на дне стека – сейчас доступно больше данных, чем когда-либо, и лишь некоторые из них агрегированы.

Единственный способ для профессионалов данных справиться с давлением неоднородности как сверху, так и снизу стека состоит в том, чтобы использовать новый подход к управлению данными. Он объединяет операции и сотрудничество для организации и доставки данных из многих источников, надежно совместимым с происхождением необходимых для поддержки воспроизводимых потоков данных.

Сегодня инфраструктура, необходимая для поддержки количества, скорости и разнообразия данных, имеющихся на предприятии, радикально отличается от того, что предполагали традиционные подходы к управлению данными. Характер DataOps

включает в себя необходимость управления многими источниками данных и многопотоковыми конвейерами данных с широким спектром преобразования.



DataOps по версии Эндрю Палмера

Звучит адски сложно, но тем не менее это факт.

Выбор подхода по организации работы с загрузкой, обработкой и агрегацией данных для разного количества данных будет зависеть от многих факторов, которые могут быть индивидуальны для организации. Например, если количество источников данных мало, и они контролируются централизованно со стороны организации, то DataOps как подход избыточен. Но если у организации множество источников данных, различные потребители, потребности в аналитических сервисах и нет возможности проконтролировать сам источник, то единственным эффективным решением по организации работы команды будет DataOps.

INDY Rock's
VS
BIG
DATA
PROFILE
ID CITIZENS
TRUTH
SECURITY
CONTRACT DATA
PERSONAL DETAILS
NEWS
GAMES
CUSTOMERS
TRUST APPS
SUBSCRIPTIONS
SERVICES MOBILE RULES
PRODUCTS PRIVATE
OWNERSHIP
FINTECH SAFE
EVALUATION
FACEBOOK
GOOGLE
FLOW
FACTORY
IOT
DATA

Глава 3
Storytelling с данными

Когда ты уже сделал большую работу, повозился с базами данных, нашел и обработал эти данные, то остается, как ни странно, самое сложное – умение их правильно показать.

Многим не составляет труда вылить на голову другому человеку результаты своего анализа. Вроде все правильно, но ощущение что тот ни черта не понял. И это очень частая проблема. Мне даже порой кажется, что эта проблема стоит выше всех остальных проблем, которые возникают при работе с данными.

По сути, ведь как – постороннему человеку должен быть понятен не только результат ночных блужданий по данным, но и то, что этот результат из себя представляет. И тут, как говорится, одного рецепта нет на всех, но я постарался структурировать лучшие практики и выделить только самое ключевое.

Итак, демонстрация результата работы с данными – один из важнейших этапов извлечения ценности из данных, который включает в себя визуализацию, описание предмета исследования и самих **данных**[52].

В одном из подходов, сформулированных известным экспертом в области данных, Брентом Дюксом[53], успешность представления данных зависит от того, насколько будет хорошо структурирован контекст в отношении той или иной аудитории.

Аудитория как таковая не воспринимает сухие цифры. Совсем. Нужно рассказать про принцессу, показать, как она убила дракона и спасла рыцаря, ну или наоборот.

В общем, нужна драма, чтобы вызвать взаимный интерес у людей, когда им все равно, что ты будешь рассказывать.

Аналитические отчеты, информационные записки или аналитическое прикрытие для этого мало эффективно. Люди хотят быть частью рассказа, поэтому при формулировании представления, демонстрации или презентации, упор нужно сделать вокруг так называемого «Aha Moment» – момента, в котором каждый из слушателей начинает воспринимать себя неотъемлемой частью рассказа.

Так, по данным исследований **нейрофизиологов**[54] [55],оказалось, что в основе эффективных решений лежит не логика, а эмоции. И это факт. Ведь есть же целое исследование.

Что такое Data Storytelling

В условиях неопределенности и волатильности на смену взвешенному интеллекту приходит эмоциональная оценка ценности того или иного события. То есть люди начинают в первую очередь переживать, когда вокруг наступает полнейший хаос. Если коротко, один из основных принципов звучит следующим образом: «Если что-то доставляет больше радости или делает вас сильно несчастными, оно будет оценено совершенно иным образом, нежели логическая цепочка рассуждений». Значит, это нужно использовать!

Истории меняют работу **головного мозга**[56]. Такие зоны как **область Вернике**[57], а также **Центр Брока**[58] вовлечены в то, как мы распознаем текст. В частности, при чтении слов возникает не просто ассоциативный ряд, но и подключаются другие регионы головного мозга, например, отвечающие за обоняние, если читатель прочел, скажем, слово «кофе» или «секс». В этом случае читатель подумал про конкретную ассоциацию, доставляющую ему удовольствие.

Итак, когда люди читают «голые» цифры, на самом деле, они

ощущают истории. Все это ведет к нескольким очень важным аспектам, таким как:

• **Запоминаемость** – в исследовании профессора Стэнфорда, Чип Хилс, обнаружил, что 63 процента слушателей помнят **истории**[59] и только пять процентов помнят цифры или сухие факты. Истории – это то, что привлекает внимание людей, остальное они просто не запоминают.

• **Убедительность** – история о помощи африканским детям показала, что сухие факты менее эффективны для сбора средств, чем история конкретного семилетнего мальчика Рокиа из Мали, которому все сопереживали.

• **Вовлеченность** – хорошая история всегда рассматривается как **подарок**[60]. Люди входят в особое состояние, когда слышат отличную историю, так что они забывают о скепсисе и фокусируются на том, чем эта история завершится.

Идеальная история отвечаем на ключевые вопросы

Идем дальше. Помимо рассказа мы должны помнить, что любой бизнес представляет из себя сложную форму кооперации людей. Всякие сложные союзы, опасные игры и прочие непростые политические моменты. Но любой сложный бизнес хочет получать ответы на регулярные и простые вопросы, которые помогают двигаться дальше. Сложный бизнес хочет получать обратную связь, чтобы бизнес-лидеры смогли понять, что именно происходит не так, почему нарушаются их ожидания.

В какой-то момент я понял, что прежде, чем придумать историю, нужно разобраться в том, что именно будет моим инструментом для хорошей истории.

Большинство вопросов, поднимаемых ежедневно, можно стандартизировать, а значит, они конечны. Как я это понял? Очень просто.

Бизнес – это всегда набор функций. Функции в основном всегда состоят из отдельных департаментов, и такое разделение существует для того, чтобы решать вполне конкретные проблемы.

Существует также ряд вопросов, которые становятся уникальными. Ответ на них представляет собой полноценный кейс.

Кейс необходим для того, чтобы организация не задавала один и тот же вопрос несколько раз. От этого сегодня страдает большинство организаций. Обычно, чтобы ответить на эти вопросы, каждый уважающий себя руководитель департамента создает под столом свой собственный отдел аналитики, так называемый Data Silos, который фактически отвечает на одни и те же вопросы без возможности скоординированной работы над ними.

Усиление коллаборации между разными Data Silos – одна из важнейших задач при организации работы аналитической функции внутри организации.

Важно помнить, что сторителлинг конкретного кейса с использованием данных отличается от сервиса, который позволяет регулярно отвечать на вопрос. Рассмотрим пример. К вам может обратиться руководитель по маркетингу в конкретном банке с запросом помочь разобраться, почему маркетинговая кампания по привлечению новых клиентов и активации мобильного банка дает крайне низкие результаты. При этом не стоит ждать детальной постановки задачи или описания данных,

с которыми вам предстоит работать. В большинстве случае разбор аналогичных кейсов – это постоянная работа с неизвестными переменными, поиск и исследование неописанных процессов, а также выявление проблематики.

В случае с кампанией, низкий результат отклика вполне может иметь ряд причин, которые вы сможете выявить в процессе исследования:

• **Некорректная выборка клиентских данных**, участвующих в рассылке уведомлений и предложений. Из-за ошибок в качестве данных вместо ожидаемых сегментов, могут быть совсем не те, или в один сегмент могут быть объединены несколько клиентских сегментов, часть из которых требует дополнительной стимуляции до момента принятия решения. Например, по ряду причин случайно объединили клиентов из возрастной группы (50+ и выше) и молодой возрастной группой (20+). Так, группа 50+ в большинстве случае будет требовать дополнительной коммуникации, а также вполне возможно, что выбранный канал коммуникации будет неэффективен для них.

• **Ошибки в контактных данных**. Человеческий фактор может сработать в случае, если в процессах нет достаточной степени контроля по работе с данными. Например, при работе с зарплатным реестром (реестр сотрудников предприятий, получаемый банками для выпуска банковских карт) сотрудник, выполняющий проверку данных на стороне предприятия, может не иметь доступа к контактным данным и для каждого из сотрудников предприятия укажет контактный номер бухгалтера предприятия. В этом случае вся коммуникация не дойдет до конкретного сотрудника.

• **Текст коммуникации и интеграция в процессы.** Текст коммуникации так же может содержать ошибки или неточности, например, вместе с предложением по активации мобильного банка может идти предложение по тому или иному банковскому продукту (от открытия вклада до оформления карты), при этом может быть непонятно, как именно оформлять этот самый продукт. Банковские отделения могут не иметь конкретного бизнес-процесса, поддерживающего оформление продуктов в заявленной кампании: при предложении активировать интернет-банк, в отделениях может не быть соответствующего процесса, где клиентские менеджеры смогут объяснить, как именно это сделать.

При построении и исследовании «воронки»[61], список проблем обычно достаточно широк и не ограничивается указанными кейсами. Таким образом, за задачей исследования «почему конверсия маркетинговой кампании находится на

низком уровне» может лежать целый пул проблем. Как их правильно скоммуницировать? Большая часть из них касается работы нескольких подразделений и может носить системных характер, то есть такие проблемы могут повторяться.

Проблемы могут быть специфичны, и их понимание может потребовать определенного погружения в то, как выглядит и работает текущий бизнес-процесс запуска новой кампании. К такому погружению заказчик исследования может быть не готов:

• Слишком мало времени доступно для подобного погружения.

• Коммуникационные барьеры – погружение потребует определенных навыков работы с данными, которыми руководитель может не обладать.

• Доверие – руководитель должен довериться в таком погружении сторонней компетенции.

Можно подготовить исследование, но если на его основе не будет принято ни одного решения или не будут решены те проблемы, которые выявлены, то, считайте, что работа выполнена впустую.

ВАШ ДЭШБОРД УМЕР

Итак, основа успешного кейса – история. Необходимо иметь план, процесс и историю. Рассмотрим историю, предложенную компанией SAS.

В США четыре из пяти наименее любимых брендов – это банкинг. Если предприниматель будет нанимать новых сотрудников, ему проще выписывать чеки и платить за это, что, в свою очередь, проще, чем идти в банк, оформлять счет и зарплатные карты, потому что выстраивать взаимоотношения с небольшим банком слишком сложно. Вот вам вопрос: «Что банк как партнер может сделать, чтобы привлечь новых клиентов?»

Известная американская компания SAS предлагает методологию из **четырех шагов**[62][63]:

• **Запуск (The setup)** – этап не обязательно предполагает выбор между чем-то трагическим или смешным. Достаточно, чтобы было интересно. В приведенном примере можно начать с вопроса: «Почему люди перестают использовать традиционные банковские сервисы?»

• **Контекст (The context)** – этап предполагает, что необходимо обрисовать общую картину и заполнить те пробелы, которые были обозначены на предыдущем этапе. Лучше использовать визуальные паттерны, чем просто текст. В

приведенном примере можно указать следующий контекст: «Отказ от банковских сервисов происходит из-за изменений в жизни, а также фрустрации по отношению к доверию, приватности и низком качеству клиентского обслуживания».

• **Опции (The options)** – все это приводит к этапу, где рассматриваются возможные решения. В приведенном примере: «Как настройки и функционал в мобильном приложении позволяют предпринимателю видеть ценность?»

• **Решение (The action)** – на этом этапе необходимо стимулировать аудиторию к действию. Действие может быть любым, главное, чтобы оно подходило по сценарию. В случае с предпринимателем – «дать простую и обратную форму для обратной связи, возможно это будет шаг к еще большей истории».

Нет необходимости готовить стандартные формы отчетов или налаживать систему уведомлений по аномалии. Сторителлинг не предполагает и не призывает аудиторию следовать за процессом, наоборот – это постпроцессное состояние, где самое главное – история.

Если вы находитесь на рабочей встрече, где большинство участников смотрят в цифры и изучают графики, и в вашей зоне компетенции работа с данными, первое что вам следует сделать – отказаться от отчетности.

Отчетность сегодня – это артефакт регулярного процесса. Есть более эффективные способы выявлять отклонения в процессе, которые могут привести к снижению создания ценности в бизнесе. Но все это элементы оперативного управления.

Переходя на стратегический и тактический уровень, нужно понимать, что не так в бизнесе на уровне клиента и учитывает ли развитие организации эти сложности, уязвимости и риски. Здесь неважны показатели конверсии каналов, проникновение продуктов на клиента или среднее время до сделки, хотя большинство компаний по-прежнему в это верит. По-настоящему, всем заправляют истории, именно на них фокусируется внимание менеджмента, когда он погряз в операционной волоките.

Ханс Рослинг – один из самых ранних идеологов сторителлинга с использованием данных, совместно с фондом Gapminder.org разработал решение по демонстрированию важнейших фактов из истории развития человечества – **Trendalyzer**[64][65].

В кратчайшие сроки, более двенадцати лет назад, он построил диалог с публикой в своем выступлении на TED и смог его визуализировать, поставив интереснейший вопрос о том, является ли деление мира на «развитые» и «развивающиеся»

страны справедливым. Используя в своем выступлении данные, собранные из различных публичных источников, таких как ООН, он явно показывает, что подобная интерпретация или классификация географических регионов более иррелевантна и имела место только в 1970 году, а с тех пор произошел гигантский скачок в росте уровня продолжительности жизни и повышения уровня благосостояния.

Впоследствии, он регулярно проводил эксперименты по анимированию данных и рассказыванию историй на основании данных. Так, в 2017 году, в одном из своих экспериментов совместно с BBC, он построил рассказ о влиянии третьей промышленной революции на развитие уровня жизни в двухстах различных странах за прошедшие **двести лет**[66], включая анализ влияния войн, эпидемий и мировых конфликтов. Эксперимент длился ровно 4 минуты и 42 секунды.

В своих рассказах он использовал только один инструмент, разработанный его командой, визуализируя тренды и интегрируя новый контент. Сложные расчеты и формулы в его рассказе превращались в контекст, который был понятен каждому из участников его лекций, даже самым неподготовленным.

Помимо инструментов, всегда следует обращать внимание на корректность применяемой методологии расчета показателей.

Например, в одном из своих выступлений, касающихся исследования снижения уровня смертности, **Ханс Рослинг**[67] предположил, что использование показателя средневзвешенного ежегодного прироста для определения размера снижения уровня смертности – некорректно и вводит читателя в заблуждение.

Использование некорректных метрик и показателей – для бизнеса не исключение. При работе с аналитикой часто теряется нить, так что в конечной презентации менеджмент показывают друг другу иррелевантные цифры, которые в принципе не могут существовать рядом. Но ввиду сложности изложения данные факты обычно незаметны для менеджмента. Если весь анализ операционной деятельности компании так же отточен в виде процесса, то никто из менеджмента не будет поднимать вопрос о неуместности тех или иных показателей.

Тем и хорош сторителлинг, он заставляет взглянуть на все с чистого листа.

С 2015 по 2017 годы порталы **Import.io**[68], **rededit**[69] и журнал **Economist**[70] собрали лучшие примеры Data Storytelling за последние два века:

• Картографическая визуализация о вторжении Наполеона в Россию, 29 ноября 1869, подготовленная Шарлем Жозефом Минаром, французским инженером, топографом и автором

проектов портов и каналов. Карта включала в себя **6 (!)** различных видов данных:

◊ География – реки, города и сражения привязаны к реальным географическим локациям, где они проходили.

◊ Путь движения армии – направление вторжения армии Наполеона в Россию.

◊ Путь отступления армии – детально проработанный путь отступления армии Наполеона после поражения.

◊ Численность войск – количество оставшихся солдат по мере движения армии (каждый миллиметр представляет десять тысяч человек). Поражает размер понесенных потерь. Наполеон вторгся в Россию с армией в 442 000 солдат, дошел до Москвы с численностью уже в 100 000 солдат и бежал из России небольшим полком в 10 000 человек.

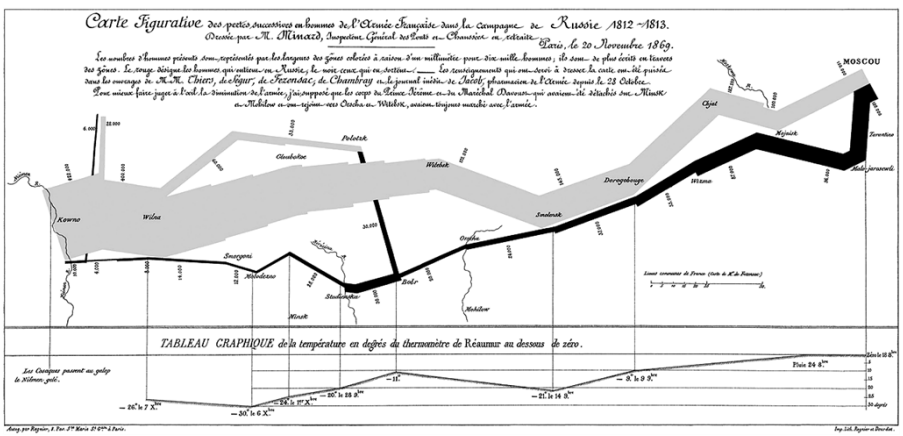


Диаграмма эффективности вторжении армии Наполеона в Россию 1812–1813

◦ Температура – в зависимости от продвижения армии, указано снижение температуры.

◦ Время – все данные соотносятся с временной шкалой.

• Круговая диаграмма о количестве смертей в Крымской войне, которую подготовила сестра милосердия и общественная деятельница Великобритании Фроленс Найтингейл. В своих трудах она впервые использовала круговые диаграммы, тем самым став их изобретателем. Она подготовила эти диаграммы, чтобы показать, сколько смертей можно было избежать, если заниматься профилактикой и лечением заболеваний раненных солдат, которые подвергались двойному риску с попаданием в госпитали. На рисунке выделен размер смертности, который наступал от болезней или инфекций уже в госпиталях, куда попадали раненые солдаты. Диаграмма сестры Найтингейл

конкретна, наглядна и имеет четкий «call to action», но она не идеальна, как утверждает журнал Economist. Так, каждый из цветных клиньев измеряется из центра, поэтому частично закрывает друг друга (вот только эта книга черно-белая, поэтому советую найти диаграмму в Интернете). В дополнении число смертей не указано, хотя это был относительный размер. Но даже несмотря на это, данная инфографика была включена в отчет комиссии по проблемам здоровья в армии, и оказала положительное воздействие на принимаемые решения.

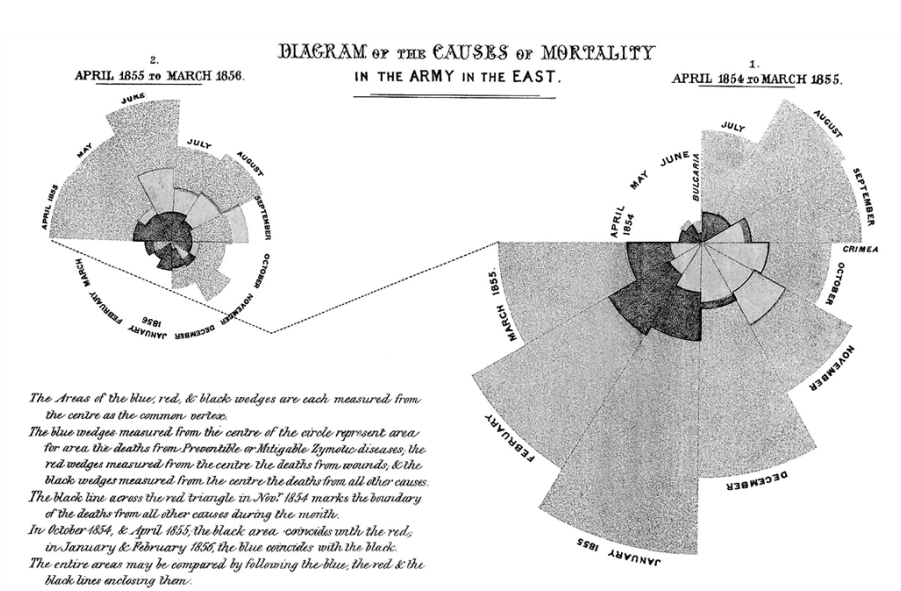


Диаграмма причин смертности в Армии на Востоке

• Экономические диаграммы и чарты шотландского инженера и основателя графических методов статистики Уильяма Плейфэра. Плейфэр изобрел линейчатый график и гистограммы для представления данных. Ряд его диаграмм отражает торговый баланс для Англии. Он был первым, кто показал размеры и экспорт на одном графике, сформулировав тезис о влиянии сдвига торгового баланса на уровень развития той или иной страны.

Один из самых известных его графиков отражает еженедельную заработную плату хорошего механика. Этим графиком он пытался пояснить связь себестоимости пшеницы и стоимости механистического труда. Один из выводов графика: стоимость пшеницы сегодня стала несоизмеримо мала с переходом к механистическому труду. Использование Плейфэром горизонтальной и вертикальной осей для представления времени и денег стало новшеством для того времени. Он был первым, кто

использовал данные не только для того, чтобы информировать, но и для того, чтобы убеждать принимать решения и проводить кампании.

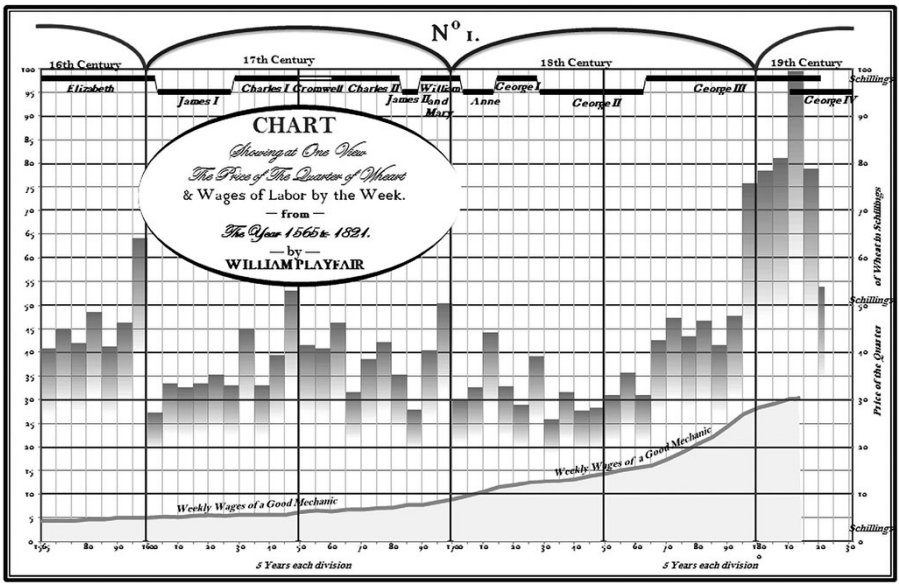
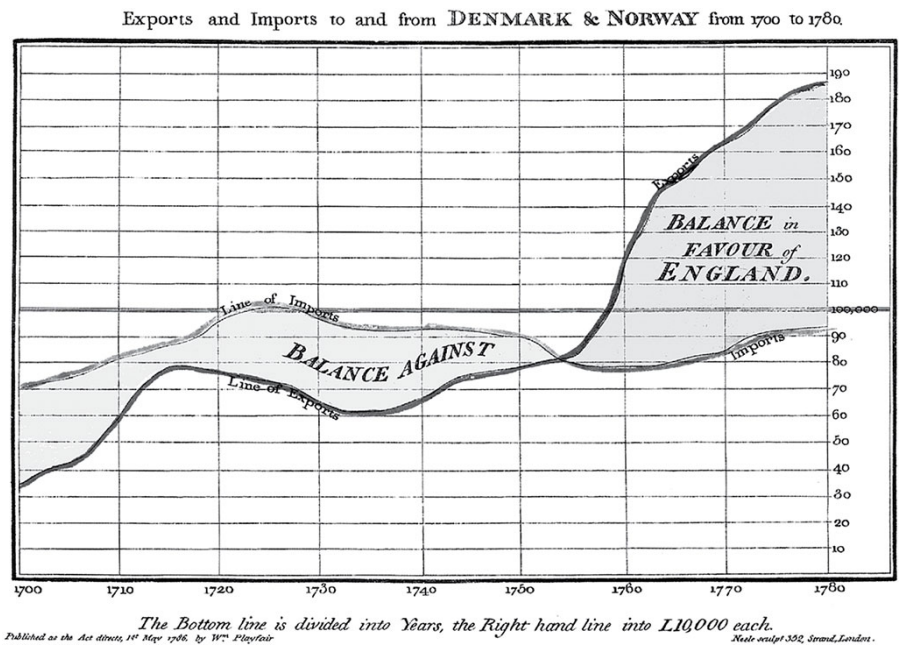


График роста заработной платы хорошего механика



Импорт и экспорт для Дании и Норвегии с 1700 по 1780

• Самые громкие районы Нью-Йорка. В своей **статье**[71] в январе 2015 в журнале The NewYorker известный аналитик,

преподаватель **Института Пратта**[72] и автор проекта I Quant NY (Я считаю NY), Бен Веллингтон, используя публичные данные, определил худшие для проживания районы Нью-Йорка по уровню шума. Проанализировав за несколько лет все обращения по уровню шума среди жителей мегаполиса, Бен категоризировал все обращения жителей по темам и по географии, определив районы с наиболее высоким уровнем шума. Самым шумным стал район Мидтауна Манхэттена, где среди лидеров раздражения были строительные работы, вечеринки, громкая музыка и громкие разговоры. Статья вызвала большой резонанс в обществе, на что Департамент полиции и Департамент защиты окружающей среды взяли на себя обязательства разработать индивидуальные решения для различных районов города. С наглядным результатом анализа Веллингтона можно ознакомиться здесь:

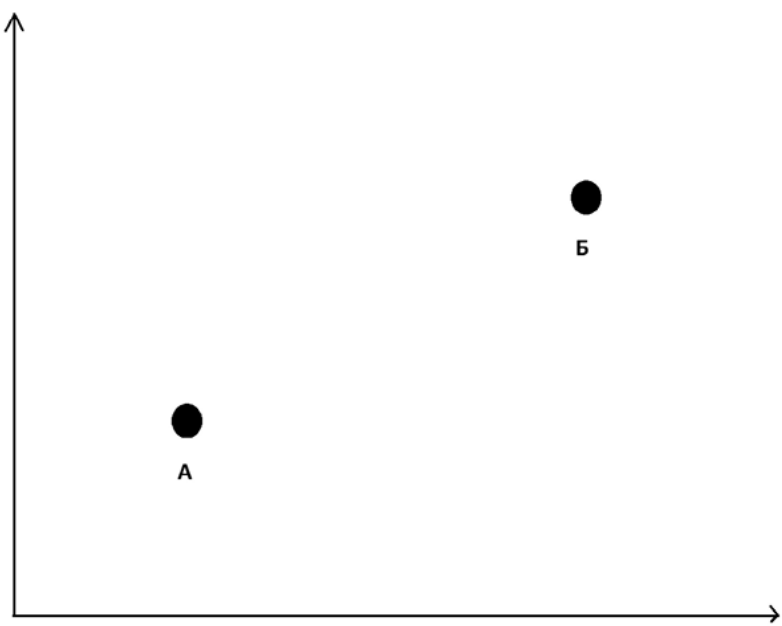
Не благодарите.

Декодирование аналитического контента требует усилий

В 1984 Уильям Кливленд и Роберт МакГил, известные исследователи в области статистики, в своей работе **«Восприятие графики»**[73] (журнал Американской Ассоциации Статистики, № 79 от 1984) выявили, что человек очень плохо интерпретирует ряд объектов и форм, если с их помощью отражается аналитическая информация. Их исследование стало одним из первых, структурирующих подход в восприятии человеком аналитической информации.

Трехмерные объекты, углы, кривые или окружности – все это крайне сложно понять, а затем еще и интерпретировать количественные данные при наблюдении за аналитическим отчетом. Выбор фреймворка и паттерна для визуализации оказывает крайне сильное влияние на возможность человека декодировать аналитический контент, который ранее был подготовлен с использованием данных.

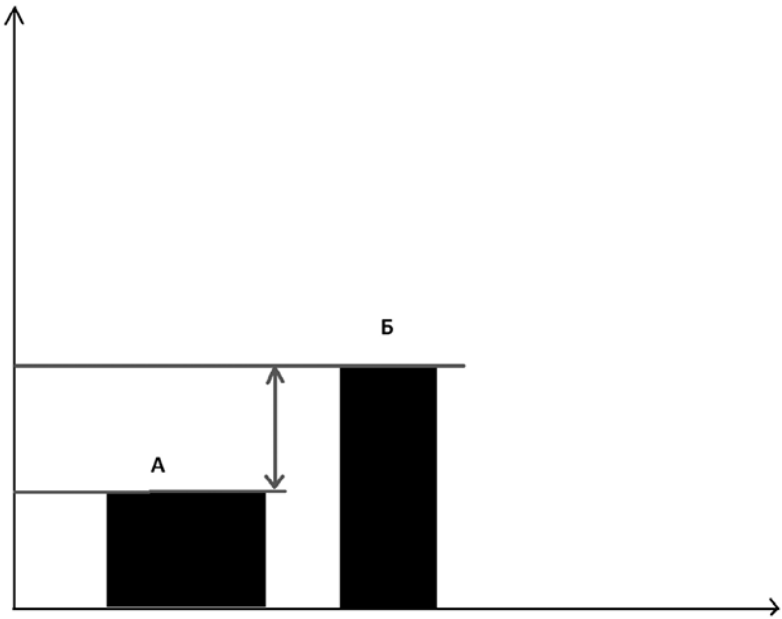
Например, какое из чисел больше? А или В? Насколько оно больше?



Восприятие большего числа – МасГил

Для большинства очевидно, что число В больше, чем А в два

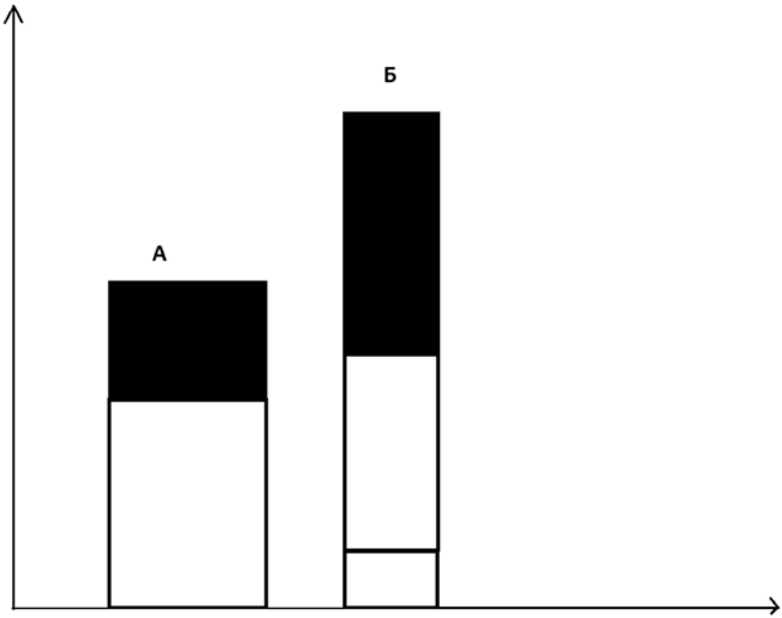
небольшим раза.



Сравнение объектов

А теперь попытайтесь быстро ответить, какое из делений больше, и как именно они соотносятся друг с другом в процентном выражении?

В своем исследовании МакГил указывает, что человек принимает решение о декодировании аналитической информации быстро, используя интуицию, без погружения в сложные расчеты.



Сложность сравнения длины делений для разной позиции

В первом случае, так как деления находятся на общем уровне, человек делает свой вывод с использованием общего уровня. Во втором случае нельзя использовать общий уровень, для сравнения размеров потребуется провести ряд аналитических расчетов для того, чтобы измерить, как именно отличается высота делений.

Кливленд и МакГил рассмотрели пять примеров чартов и провели исследование с привлечением студентов и преподавателей соответствующих направлений. Все собранные ответы они разделили на правильные и неправильные и измерили размер допущенной ошибки в зависимости от того, как именно располагались сравниваемые деления чартов по отношению друг к другу на каждом из пять чартов.

Оказалось, что чем ближе друг к другу сравниваемые деления, тем выше точность декодирования аналитического контекста со стороны человека, а чем деления дальше друг от друга, тем вероятнее рост ошибки. Когнитивное восприятие имеет свой заданный шаблон в зависимости от типа используемых чартов. Для чартов, где находятся деления, которые нужно сравнить, человек по умолчанию ищет сравнение в отношении общей линии или позиции. Если человек видит карту, то включается шаблон анализа насыщенности цветом, который используется на карте.

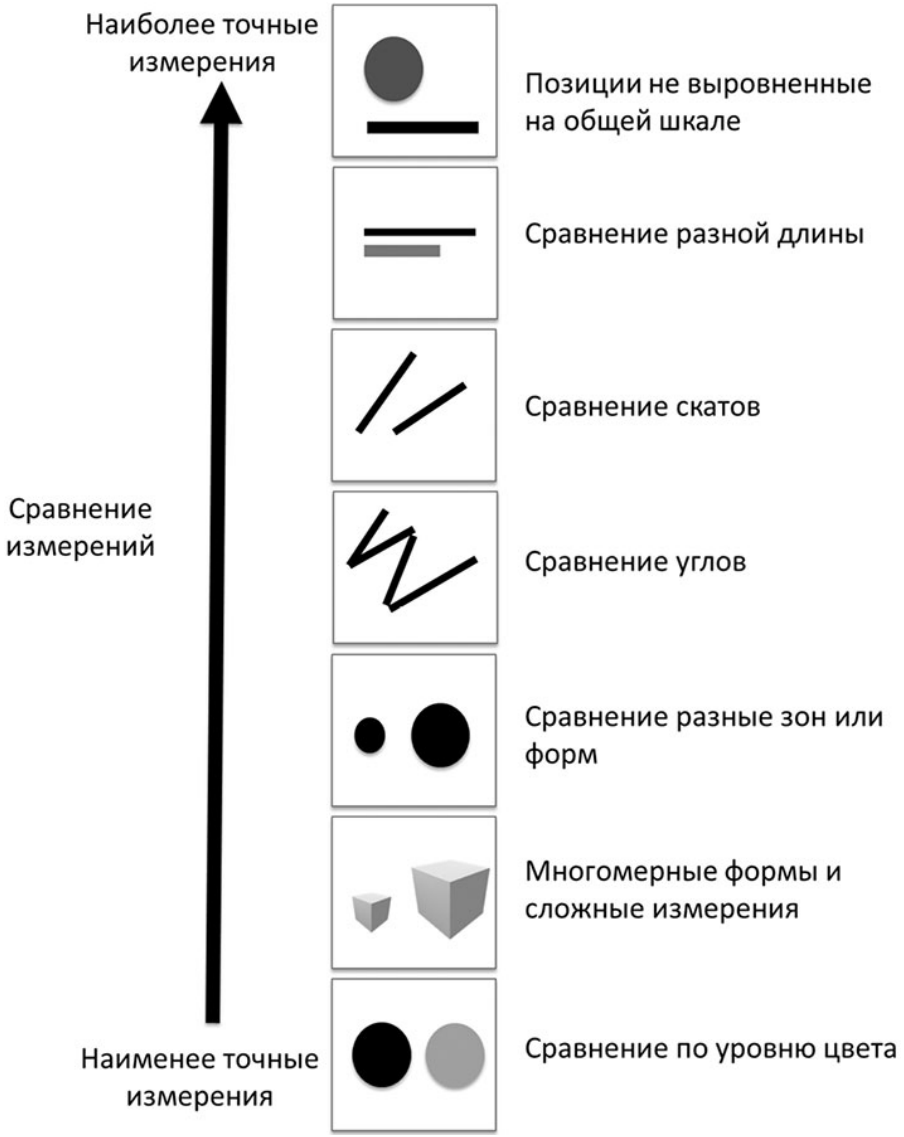
Продолжая эксперимент профессора Стэнфорда, Джеффри Хиир и Майкл Бостоk, используя анализ и результаты МакГил

и Кливленд, выявили, что круговые диаграммы – наиболее сложный объект по интерпретации, и их восприятие пользователями уже несет в себе ряд ошибок. Отчасти именно поэтому ни одно существенное научное исследование сегодня не использует круговые диаграммы в описании результатов работы, так как их восприятие сильно разнится между разными категориями пользователей. Это относится и к восприятию геометрических объектов (сравнение углов и зон), а значит использование круговых диаграмм и вовсе искажает аналитический контекст, подаваемый слушателю.

Тем самым, научное сообщество сформулировало фреймворк по восприятию аналитических данных, который популярен и по сей день. Разделяя по сложности интерпретации аналитических материалов, мы имеем следующее:

- Сравнение на общем уровне
- Сравнение объектов не на общем уровне
- Сравнение длины, угла или направления
- Сравнение зон
- Сравнение объемов или размеров
- Сравнение цвета (насыщенность и так далее)

Чем выше сложность распознавания, тем выше вероятность ошибки или искажения, с которым пользователи будут воспринимать **контекст**[74].[75]



Ранжирование визуальных кодировок по точности восприятия

Большинство решений представляют собой конечное решение, которое не учитывает в себе эти особенности восприятия. Впоследствии Джефри Хиир и Макcл Босток разработали ряд библиотек и фреймворков для визуализации данных, которые учитывают эти зависимости по сложности восприятия.

• **Protovis**[76] – библиотека в JavaScript, которая позволяет управлять внешним видом графика через скрипт с определенным синтаксисом и использует **Canvas**[77] чтобы бы можно было интегрировать графики в веб-страницы, делать их красивыми,

многоуровневыми и интегрировать видео или иной активный контент прямо в аналитику.

• **Flare**[78] – фреймворк на python, который позволяет быстро строить модели и взаимосвязи.

• **Vega**[79] – формат данных, который позволяет сохранять и управлять чартами, графиками и аналитикой, в том числе с возможностью воспроизводить их в браузере, поддерживающем HTML5.

• **D3 Data-Driven Documents** – библиотека для JavaScript, используемая веб-сайтами, которая позволяет анализировать и работать с данными используя браузер.

Визуальная часть, как отмечалось ранее, одна из трех основных частей, участвующих при демонстрации конечных результатов и формировании рассказа. Успешность восприятия или декодирования аналитической информации тесно связана с тем, как эта информация представлена.

В 1982 году Эдвард Тафт, американский статистик, профессор статистики, политологии и компьютерных наук Йельского университета, сформулировал и опубликовал ряд важнейших принципов в графическом дизайне в книге «Visual Display of Quantitative Information»:

• Использовать историю для пояснения описания данных.

• Тщательно выбирать формат представления.

• Интегрировать описание текста с изображениями.

• Отражать и сравнивать объекты соответственно их размеру.

• Избегать использование декоративного контента.

В процессе своих исследований позднее Эдвард Тафт также придумал новую форму транслирования аналитического контента – микрочарты (искрографики) или спарклайны. Это небольшие микрочарты размером в несколько слов, отражающие какую-то определенную динамику или отвечающие на какой-то конкретный вопрос. Функциональность микрочартов была впоследствии применена практически в большинстве аналитических средств, и одним из самых массовых применений стал Microsoft Excel.

Таким образом, ошибки в выборе визуализации крайне серьезно влияют на конечное восприятие доклада или отчета, когда он представляется широкой публике.

Impact investment – у каждого рассказа должна быть цель

Финальный блок успешного рассказа с использованием данных – это фокус на влиянии, которое он способен оказать. Любая инвестиция времени, посвященная исследованию и анализу данных, должна приводить к формированию конкретного результата.

В 2012 году мэр Нью-Йорка Майкл Блумберг подписал специальный закон «Open Data Law», обязывающий городские власти раскрывать свои данные для пользования, открыв тем самым целую новую главу создания совместных сервисов с использованием данных. Закон установил порядок раскрытия и перечень информации, которую обязаны были предоставлять власти с максимальным сроком раскрытия не позднее 31 декабря 2018. Раскрытие данных происходило неравномерно, власти раскрывали свои данные постепенно, поэтому, чтобы получить интересующие данные, необходимо было заполнять специальную форму запроса (FOIL FORM) для того, чтобы департамент той или иной службы предоставил запрашиваемые данные. Существенная часть данных раскрывалась в PDF-файлах, затрудняя их обработку и анализ. Например, данные по ДТП публиковались только в PDF, составляя сотни и тысячи документов. Так продолжалось, пока один из разработчиков по имени **Джон Краусс**[80][81] не придумал собственную программу для конвертации файлов PDF в CSV, чтобы их можно было уже загрузить в аналитические средства. Сообщества неоднократно в своих выступлениях делали акцент на трансформации используемого формата предоставления информации, пока администрация де Блазио[82] не пересмотрела интерфейсы предоставления данных.

Бен Веллингтон основал свой проект IQuantNY и стал использовать эти данные, чтобы повлиять на политику властей Нью-Йорка. Каждое выступление или пост в своем блоге он посвящал конкретным проблемам, призывая власти обратить внимание. В какой-то момент он добился определенного результата в этом направлении. Вот несколько наиболее ярких исследований и публикаций, которые он сделал.

• Ошибка городского бюджета на 791 миллион долларов – в 2016 году Администрация наконец опубликовала городской бюджет на 2017 год на портале Открытых данных. При детальном анализе Бен выявил ошибку в 791 миллион долларов по статье

финансирования Департамента полиции Нью-Йорка в части защиты иностранных представительств. Официальный ответ городской администрации указал, что, действительно, это была ошибка в аллокации средств. Корректное значение аллокации средств на 2017 год составляло не более 25 миллионов долларов.

• Штрафы за оплаченные парковки – в Нью-Йорке парковать автомобиль можно было только на специально отведенных местах. В 2009 году Администрация внесла изменения в правила парковки и разрешила оставлять автомобили рядом с пожарными гидрантами, возле которых было свободное место. В своем исследовании Бен обнаружил серию регулярно выдаваемых на протяжении нескольких лет штрафов в местах, где стояли гидранты, но не было запрещающей стоянку разметки. Иными словами, огромное количество штрафов на сумму более 1,7 миллиона долларов было выписано нелегально, так как автомобиль находился в разрешенной зоне парковки. Таких зон было выявлено порядка 1966. Происходило это потому, что большая часть сотрудников полиции проигнорировала изменения законодательства в 2009 году. Администрация признала ошибку, допущенную патрульными службами, сформировав дополнительный фокус на переобучение патрульных служб. Вскоре была проведена корректировка разметки во всех выявленных местах.

• Самая грязная вода в Нью-Йорке – сведение статистики по самым грязным водоемам в городской черте. На портале открытых данных Нью-Йорка находился реестр с анализами данных по водоемам за несколько месяцев. Для проведения подробного анализа понадобился полный массив данных, который находился на отдельном сайте Департамента охраны окружающей среды. Данные были разбиты на много раздельных Excel-реестров с различными заголовками, которые нужно было свести вместе. Уровень загрязнений существенно превышал норму, в самых грязных районах превышение нормы было многократным. С вероятностью в 94 процента купание в водоеме могло привести человека к летальному исходу. Одним из таких мест оказалось Coney Island Creek. В результате, Бен опять привлек внимание Администрации и Департамента защиты окружающей среды. Он выписал ряд крупнейших штрафов по 400 тысяч долларов комплексу апартаментов, находящемуся в зоне загрязнения, большая часть из которых была направлена в Фонд защиты дикой природы. Тем не менее сообщество разделилось, требуя увеличить размер штрафов в десятки раз, доведя его до

четырех миллионов долларов, аргументируя это тем, что уровень ущерба от загрязнения выше, чем размер штрафов.

Итак, каждая история – это большая проделанная работа по обработке, анализу, гармонизации и нормализации данных. В процессе выполнения сложной и рутинной работы всегда снижаются ожидания от аудитории, которая думает, что ничего важного не произойдет после демонстрации результатов. Но это не так.

Любые изменения происходят медленно, но они происходят, если есть для этого стимул. Выявить этот стимул и отразить в своей работе – ключевая задача Data Journalizm.

Глава 4
Регулирование данных

Р – регулирование.

Данные и капитализм.

Штука бесполезная и беспощадная во всех отношениях. Ценности в регулировании де-факто мало, оно лишь снижает скорость развития в цифровой экономике.

Нет однозначной истории, как же регулировать данные.

Во-первых, с юридической точки зрения надо определить, что такое данные. А с этим не только в нашей стране беда, но и в международном пространстве нет единства и понимания по таким вопросам.

В США, например, нет законов, прямо регулирующих Большие данные. Иными словами, ты можешь пострадать, только если косвенно затронешь чьи-то интересы, и они это докажут.

В самом начале я говорил, что есть две стратегии работы с данными:

1. Либо нападение – то есть используем те данные, которые есть, с целью побольше заработать.

2. Либо защита – сидим на данных, никому не даем и защищаемся от всячески возможных рисков и возникающих сложностей.

Итак, родина капитализма, конечно же, предпочла первый вариант. Что там делают с вашими данными – похоже на какофонию и безумную спонтанную оргию организаций, которых вместе никто не собирал и к сотрудничеству не приглашал.

Конечно, есть небольшие исключения, и это хорошо. Они как раз структурируют и задают общий тон того, что делать можно, а что – не очень.

Кстати, именно в США в 2010 году был известный скандал в магазинах **Target**[83], когда их точнейшие аналитические алгоритмы определили, что несовершеннолетняя школьница ждет ребенка. Конечно, первым прибежал ее отец и чуть не поубивал менеджеров Target за непристойный контент и предложения школьнице рожать.

А все началось с Эндрю Пола, который в 2002 году работал статистиком в Target. К нему подошли его коллеги и спросили его, «может ли он выяснить, беременный покупатель или нет, даже он не хочет, чтобы мы как магазин знали?».

Эндрю имел магистерскую степень по статистике и еще одну – по экономике, и, конечно же, был повернут на анализе

поведения пользователей с использованием данных.

Спустя 16 лет с того безумного вопроса, ответ на который сделал Эндрю Пола мегазнаменитым и успешным гостем выпусков новостей и ток-шоу в связи с эпичным скандалом, он ушел из ритейла в банкинг. Он вышел работать вице-президентом по персонализированной аналитике в пятый по величине банк США, USBank. Чем-то похоже на наш топ-5 банк, за одним исключением, что USBank ровно в десять раз больше, чем банк, находящийся на 5 месте в РФ, и даже больше, чем известный банк находящийся на первой строчке рейтинга в России. В USBank Эндрю будет развивать совершенно новое направление. Кстати, Target – это пятый по величине ритейлер в США, так что тенденция у Эндрю на лицо.

Но вернемся назад, в начало 2000-х. Сакральная мысль позади идеи взлома и анализа поведения беременных покупателей была крайне простой – молодые родители, считались Священным Граалем для сети Target. Обычно покупатели не берут все в одном магазине, они покупают везде понемногу. При этом, сеть Target продавала все виды товаров: от питания до мебели.

И, конечно, их главная мысль была стать брендом первого выбора, чтобы за любым предметом люди шли в Target. Абсолютно такая же крамольная идея лежит и в головах банкиров – стать банком первого выбора, чтобы везде платили только их карточкой.

Как мыслили тогда маркетологи: обычно человек в рутине, и до него сложно достучаться. Есть только несколько моментов в жизни, когда эта обыденность отступает, – покупатель прислушивается и готов покупать все подряд. Один из таких моментов – рождение ребенка. Тогда родители готовы перевернуть магазин вверх дном, чтобы найти самую крутую колясочку и самую прикольную кроватку. Другие события, когда человек выныривает из рутины, – например, когда слышит любимую музыку.

И тут великие маркетологи сказали Эндрю, что важен момент. А именно, момент до рождения, который потом будет занесен во все публичные источники и базы данных. Нужно ловить пап и мам тогда, когда их чадо еще не увидело свет. Было бы идеально, если бы Эндрю смог разработать модель, позволяющую вычислять второй триместр беременности, чтобы приклеить к себе покупателя на годы.

Данные, которые собирались, привязывались к общему идентификатору программы лояльности. Собственно, как и в других магазинах в любой стране мира. Идентификатор

карточки лояльности обычно связан с простеньким соцдемом – возраст, пол, есть ли дети и так далее. Дальше, используя номер телефона, например, можно взять данные из базы резюме и прикинуть, сколько человек получает на той или иной позиции. Для этого существует целая тонна обзоров и прочих сервисов.

Напомню, телефонный номер, по которому можно получить эту информацию из баз данных HeadHunter и других ресурсов, не является объектом закона о персональных данных в России.

Таким образом, мегамоллы могут быстро прикинуть, какой у вас доход, а данные операторов, помогут рассчитать, как далеко вам нужно добираться до дома. А дальше происходит простая магия чисел и вычислений. В основу модели Эндрю легли 25 продуктов, которые будущие мамы покупали в сети. «Большой бум» произошел в том, как люди пытаются предсказать поведение других людей.

Люди смогли предсказать, что нужно делать, чтобы **перестать кусать ногти**[84], или почему одни любят ходить на работу, а другие прокрастинируют каждое утро. Причем те, кто прокрастинируют, как раз и **грызут свои ногти**[85][86].

Есть в этом что-то позитивное, когда пытаешься предсказать поведение пользователей. Ищешь новые знания и все-такое. Конечно, доступ к данным тут является лютым и самым важным обстоятельством. Нейрофизиология вообще как область науки, важна для ответа на вопрос «почему мы действуем сегодня так, а не иначе».

Для получения данных по медицинским исследованиям придется разобраться в структуре закона о «Страховании здоровья и ответственности»[87] (HIPPA), который был принят в 1996 году Конгрессом и объединил регулирование как в отношении сотрудников, так и в отношении предоставления медицинских сервисов широким слоям населения. Идея HIPPA проста – стимулировать появление медицинских планов для всех групп пациентов. Основой, на которой предполагалось выполнять такое стимулирование, стал электронный документооборот истории болезни.

Спустя двадцать лет, конечно же, возникли проблемы, так как все перешло в цифровую среду. Например, те же фитнес-трекеры, которые собирают медицинскую информацию о сердцебиении, и GPS-координаты – должны соответствовать требованиям конфиденциальности медицинских данных по итогам пересмотра **HIPAA в HITECH Act**[88] в 2009 году. Это, кстати, единственное, чему они должны соответствовать сегодня в США.

В законе нет общих правил нарушения, каждый кейс должен рассматриваться индивидуально. Есть только общий штраф за

нарушения в размере 250 тысяч долларов, если это произошло в первый раз, и полутора миллиона долларов, если нарушение использования данных продолжается. Единственное, закон не так сильно соблюдается и мониторится со стороны властей.

А чтобы можно было работать с данными о финансовых транзакциях, Соединенные Штаты предлагают действовать в рамках требований по конфиденциальности в части Грэмм-Лич-Блайли акта. Этот закон инициировали республиканец Фил Грэмм и конгрессмены Джим Лич и Том Блайли. Потому что там участвовала республиканская партия, я думаю, понятно, что закон направлен на снятие ключевых барьеров в развитии банковского сектора.

Конечно, любая такая идея всегда упирается в деньги. Всегда. Просто потому, что деньги были и пока остаются единственным интерфейсом, как люди одной профессии ставят людям другой профессии разные задачки. Так, когда Эндрю Полу поставили задачку, то он представил своего виртуального покупателя, девушку, которая тратит в год не более тысячи долларов. Пол прикинул, что будет, если, например, она выйдет замуж и у нее появятся дети, увеличится ли сумма ее ежегодных трат до пяти тысяч **долларов**[89]**?** А если сравнить ее с неким виртуальным клиентом, мужчиной, который продолжает по-прежнему тратить только тысячу долларов, то логичнее было бы для Target не тратить деньги на рекламу, сегментирование таких низкодоходных клиентов и выпуск специальных предложений вроде купонов на покупку. При правильной игре такая модель, как уже понятно, позволит увеличить доход на клиента в пять раз. Вдумайтесь, в пять раз.

Чтобы построить такую модель, которая дает точность в **86,5 процента**[90], потребовалось добыть конкретные данные. Сама модель при этом оставалась возобновляемой, и ее можно было воспроизвести в разных сегментах розничной торговли, банкинге или финтехе. Ключевыми данными тут выступили:

• идентификатор карты лояльности в сети Target,

• имя и адрес проживания,

• банковская карта,

• история покупок в магазине,

• история онлайн покупок,

• номер мобильного телефона,

• действия, предпринятые пользователем в ответ на электронные письма Target в почтовом ящике (клики, переходы и так далее),

• куки и активность по поисковым запросам и просмотрам в Интернете.

Полное видео с выступления 2010 года Эндрю Пола можно увидеть **по ссылке**[91]. Позади таких исследований всегда стоит основная гипотеза, которая постепенно превратилась в аксиому «что есть паттерны поведения». Как только мозг сталкивается с определенной ситуацией, в которой он уже был, он автоматически включает определенный паттерн. Изучив такие паттерны, можно управлять знанием, экономикой и покупательской способностью. Поэтому данные будут собирать, поэтому данные будут покупать. Это гонка, и она будет только ускоряться.

Стоит отметить, что это лишь малая часть данных, которые мы оставляем о себе, и не все из них попадают под регулирование. Так, номер мобильного телефона сам по себе не является объектом регулирования для закона о персональных данных в РФ.

Суровые европейские консерваторы

Гонку сбора данных и их использования, которая началась на другом континенте, в старой доброй Европе остановили с помощью бетонной стены, которая называется GDPR – основной закон по защите данных, состоящий из 99-ти глав (!). Он вступил в силу 25 мая 2018 года. Именно в эту бетонную стену врезалось большинство продавцов данных.

Не так давно я посетил один чешский банк, и оказалось, что он не может просто так купить данные у кредитных бюро (чем он занимался постоянно) и провести скоринг потенциальных клиентов на возможность выдачи займа.

28 января 2016 года, 47 стран участников ЕС праздновали десятую годовщину Европейского Дня Защиты Данных. Этот небольшой праздник посвящен подписанной в Страсбурге членами ЕС 28 января 1981 года конвенции № 108, посвященной правилам и требованиям обработки персональных данных. Этот небольшой по объему документ вводит основные понятия: что такое персональные данные, что такое обработка файлов, что такое контроль обработки файла.

И это все, представьте себе, было больше 37-ми лет назад, когда еще и Интернета толком не было. Конвенция также вводила специальные определения данных, например, в отношении данных по здоровью, где запрещала их обработку в случае, если страна-участник ЕС не гарантирует их конфиденциальности. Саму конвенцию, в итоге, ратифицировали (в том числе и за пределами ЕС) такие страны как Мексика, Марокко, Кабо Верде, Маврикий. Огромное количество наблюдателей из других стран также сформировало свою позицию (Австралия, Япония, Корея, Израиль и другие). Сегодня Конвенция уже обновлена и актуализирована в духе времени.

Первая директива по защите данных появилась впоследствии в ЕС, в 1995 году, и была введена 13 декабря 1995 года с обязательным периодом старта применения участниками ЕС с 24 октября 1998 года. Это был достаточно масштабный по своим меркам документ, подробно описывающий как категории данных, так и категории потребителей, а еще методы защиты, арбитража работы с данными и организацию службы специальных контроллеров, которые проверяют применение директивы в каждой из стран ЕС.

Как это обычно бывает, контроль сам по себе и иезуитские методы, которые он подразумевает, крайне сложно масштабируются и спускаются на уровень конкретных

исполнителей. Отчасти поэтому первого декабря 2009 года была выпущена специальный дискуссионная бумага ЕС о будущем – «Future of Privacy».

Глобальный посыл был такой – текущее регулирование опять не работает, его нужно усложнить и обновить различными инновационными практиками, такими как «privacy by design» (приватность через дизайн) и другими.

ЕС – это сложная экосистема. Если какой-то веб-сайт обрабатывает данные резидентов/нерезидентов и находится за периметрами ЕС, то он не подпадает под регулирование Директивы 1995 года. Дискуссионная бумага подняла обсуждение о создании единого стандарта для приватности и управления данными, чтобы другие страны могли управлять своими требованиям к данным, руководствуясь общими принципами. По аналогичной схеме пошли в свое время (в далеком 1971 году) и бухгалтера, и финансисты, когда основали IASB, – специальное общество, которое впоследствии выпустило Международные стандарты финансовой отчетности (МСФО). МСФО сегодня – основа для любой финансовой публичной отчетности.

Директива 1995 года не предусматривала описания подходов, которые бы рассказали, как нужно брать с пользователей согласие на обработку данных. Поэтому нужен новый подход по работе с согласием пользователей. В общем, было много и других различных деталей, смысл обсуждения которых сводился опять же к тому, что регулирование не работает должным образом. Все это вылилось в то, что четвертого ноября 2010 года ЕС определил единую новую стратегию по защите персональных данных и начал работу по ее реализации.

Дальше, если не вдаваться в подробности, был большой спор между ЕС и США, где США хотели сохранить интерес своих компаний на территории ЕС, так что текст проекта закона менялся несколько раз крайне существенно. Пока все вокруг пытались разобрать, сколько же они тратят на комплаенс по защите данных (87 % опрошенных бизнес-субъектов не смогли определить точно цифру затрат на выполнение требований текущего и будущего законодательства), 12 марта 2014 года Европарламент общим большинством (621 за, 10 против и 22 воздержалось) утвердил реализацию GDPR, обязательную для участников ЕС с 2018 года.

Регулирует GDPR специальный офис (Informational Commisioners Office) – не путать с брутальным термином ICO, ввергающим в панику большинство традиционных банкиров. Это совсем другое.

ICO каждой страны участника ЕС входит в единый совет –

European Data Protection Board.

GDPR говорит, что все публичные компании обязаны назначить специальную роль Data Protection Officer (DPO). Такие парни стали обязательными для организаций, в которых работает более 250-ти человек и обрабатываются более пяти тысяч объектов персональных данных в год.

У любой организации есть всего лишь 72 часа на то, чтобы сообщить об утечке или нарушении GDPR в организации в случае их возникновения. Правило 72-ух часов стало новым для всех участников.

Что именно оно означает?

В течение этого срока организация, которая допустила нарушение GDPR, обязана за 72 часа (три дня) провести внутреннее расследование, информировать об этом регулятора и всех затронутых владельцев данных, а также подготовить план возмещения и купирования возникающих убытков (рисков).

Исследование в обязательном порядке должно содержать информацию о том, кто и когда имел доступ к данным, как и какие данные пользователей оказались под нарушением, каким образом эти данные были использованы.

Как человек, который занимался данными достаточно долго, с уверенностью могу сказать, что львиную часть работы в этом случае будут возлагать на директора по данным, ибо кто-то должен наладить необходимый сбор данных для регулятора.

Скорость, с которой теперь должны реагировать организации, становится феноменальной. Она соизмерима со скоростью работы комплаенса банковского сектора.

GDPR собирает лучшие практики и требования к работе с данными, в особенности к публичным компаниям.

Все проблемы и риски в отношении GDPR необходимо рассматривать на уровне Правления организации. Теперь все сделки по слияниям и поглощениям будут в обязательном порядке начинаться с анализа соответствия GDPR, так как штрафы за нарушения достаточно внушительны, вплоть от десяти миллионов долларов и до четырех процентов от годовой выручки. Неудивительно, что такая надгробная плита притормозила и припарковала развитие множества цифровых сервисов в ЕС.

«Compliance First!»

Увидеть мир в частице песка

В совокупности регулирование GDPR стало самым продвинутым с точки зрения глобального контекста. Но я все равно бы разделил мир на страны, где с точки зрения регулирования проще идти со стратегией «развития и атаки» (монетизации данных) – США, Индия, Япония, Южная Корея,

Африка, Канада и страны, где преобладает стратегия «защиты» (снижения возможных рисков и потерь) – Европа, Россия, Китай, и другие.

Различное регулирование дает различную корпоративную структуру.

Отдельно стоит отметить Великобританию, которая некогда была своего рода оазисом в части принципов корпоративного управления, задавая стандарты с помощью «Combined Code», обязательного к применению всеми публичными компаниями, размещенными на Лондонской бирже. Сегодня «Combined Code» абсолютно ничего не говорит в отношении данных, этот вакуум как раз был призван решить GDPR, но в связи с выходом Великобритании из ЕС, могу сказать, что там с данными можно творить все, что угодно, как, например, в Индии.

Крупнейший в истории человечества взлом данных, как оказалось, произошел именно в **Индии**[92]. Из платформы Aadhaar, которую приводят всем как образец биометрической системы, собирающей отпечатки пальцев, были украдены практически все записи пользователей – более одного миллиарда.

Идея создания платформы Aadhaar появилась в начале 2000-х из-за такой проблемы, как осуществление выплат социальных пособий населению Индии, не имеющему документов (паспортов). В 2009 году Правительство Индии создало единую идентификационную карту для жителей Индии. Дальше дело было за малым – привязать каждого жителя Индии к его идентификатору. И здесь самым верным решением оказался отпечаток пальца. Тогда была придумана Adhaar. За четыре года сорок процентов населения были записаны на новую платформу. А девятого октября 2013 года Национальная Платежная система Индии запустила поддержку платежных сервисов, привязанных к Aadhaar счетам для небольших сумм (не более семидесяти долларов). По итогам 2018 года Aadhaar – крупнейшая в мире база данных биометрических записей (отпечатков). Кстати, запуск Единой Биометрической Системы в России начался с анализа кейса Индии по созданию аналогичного сервиса.

Но вернемся к нашим проблемам. В 2018 году совокупно было украдено порядка двух миллиардов записей из разных источников.

Вторую строчку после Aadhaar занимает проект Exactis, который выступает брокером данных для осуществления маркетинговых компаний.

На третьем месте находится компания Under Armour, которая выявила утечку данных более 150-ти миллионов пользователей, где помимо их персональных данных (имени, пароля,

водительских прав и так далее), были еще и их медицинские данные – сердечные ритмы, нагрузки, калории и прочие данные, что каждый из нас заносит в фитнес-трекер.

Следующим лидером вслед за введением GDPR стал Китай. Пекин ввел обязательное требование держать данные о всех своих жителях на территории Китая, чтобы они стали частью единой платформы по социальному скорингу. Кто смотрел сериал «Черное Зеркало» сразу поймет, о чем речь, кто нет, – поясню. Поступки каждого жителя Китая теперь влияют на его жизнь, на то, какую работу он сможет выбрать, в какие университеты поступить, какие услуги получить от общества, какой дом купить, влияние распространяется даже на то, какой билет на высокоскоростной поезд может себе позволить человек.

Платформа социального скоринга впервые была предложена в **2014 году**[93], когда Пекин обозначил, что на основании данных будет рассчитываться единый рейтинг жителя, а в 2020-м эта схема станет обязательной для всех жителей Поднебесной, всех 1,4 миллиарда человек. Это самая настоящая система тотального контроля за гражданами, где следующий шаг после выявления нарушения – наказание. А наказывать жителей Китая могут за то, что, например, они жульничали в компьютерных играх или забронировали столик в ресторане, но не пришли. Для поддержки и обеспечения работы этой платформы Пекин разрабатывает поправки в регулировании на базе существующего законодательства E-Commerce Law, драфт которого был разослан ключевым игрокам рынка в начале 2018 года.

Новый закон четко выделяет три блока применения регулирования:

• Сетевые операторы

• Критическая информационная инфраструктура

• Провайдеры сетевых продуктов и услуг

В каждый из трех блоков попадает, к примеру, компания WeChat, – один из крупнейших в мире мессенджеров. WeChat представляет сегодня экосистему, позволяющую оплачивать покупки через qr-коды даже на овощных рынках. Qr-коды заменяют даже визитки, которые жители Китая могут дать друг другу при знакомстве. Китайское общество в своем роде перепрыгнуло эру пользования банковскими картами, так как покрытие финансовыми сервисами оставляло желать лучшего, а телефоны стали мегадоступными. Результат не заставил себя ждать.

Каждый житель Поднебесной живет и пользуется в своей бытовой жизни исключительно такими сервисами как WeChat. К слову, стоит отметить, что большая часть продуктов WeChat

находилась вне правовых рамок или попросту была незаконна до тех пор, пока не стала поистине массовой. Властям ничего другого не осталось, кроме как легализовать использование цифровых сервисов таких компаний как WeChat, Baidu и Alibaba.

Новое регулирование теперь наложит на эти компании определенные обязательства, в том числе и штрафы. Но по сравнению с Европой они, конечно, меньше (до трехсот тысяч долларов).

Россия – страна других взглядов.

Если в США известный всем парень Дуэйн «Скала» Джонсон запускает свою собственную линию одежды и кроссовки от компании Under Armor, то их раскупают менее чем за двадцать часов в день релиза коллекции. Вот это скорость! Сумасшедшая. Раскупают все подчистую, так, что кроссовок таких просто нет. Ни интернете, ни в магазинах. Не то что в РФ. В фирменных магазинах Under Armor в Москве чудо-кроссовки стоят нетронутыми, и никто их не берет. На них даже сделали сезонную скидку в двадцать процентов. Вот тебе бизнес: покупай в России неликвид и продавай в США изделия, которые делают в Китае.

Продавцы говорят, что люди интересуются и не покупают, потому что пятка в этих чудо-кроссовках проминается, а это риск. Общество у нас и так всегда под риском, и на новый еще не готово.

К чему это все? А к тому, что поведение пользователей разных стран сильно отличается, и законы тут вовсе не причем. Законы, скорее, помогают наводить порядок и определять те границы, которые сильно размыты.

Так уж повелось, что по ряду вопросов мне пришлось консультироваться с Администрацией Президента. Мне стало интересно мнение коллег и экспертов от АП, понять, как они относятся к законам. Признаться, мой собеседник оказался крайне образованным и эрудированным человеком, и он натолкнул меня на мысль о том, как должны строиться законы.

Право, которое строится на традициях, поддерживает и раскрывает взаимоотношения общественных институтов и самого общества. Под традициями каждый может понять много чего, поэтому я, конечно, тут рассматриваю все в плоскости сугубо цифровой. Что можно делать там, где еще никакой этики или традиций не существует. А что делать нельзя, даже если технология позволяет? Это пока не определено. Поэтому все существующие законы просто не раскрывают суть цифровых взаимоотношений.

Мой хороший друг, эксперт в области данных, управляющий директор по инновациям в Газпромбанке, Екатерина

Фроловичева, на открытом рабочем столе Аналитического Центра по регулированию данных в сентябре 2018 года сформулировала, что данные – это валюта.

Другие эксперты сказали, что данные – это актив. Мне нравится эта формулировка, потому что в основе любого взаимодействия в цифровой среде должна лежать сделка. Сделка по использованию данных любого человека или организации третьим лицом.

В РФ есть закон, например, № 112 от 7 июня 2013 года, который вносит изменения в основной закон, определяющий порядок доступа к информации о деятельности государственных органов и органов местного самоуправления. Доступ к данным определили, а что такое данные – нет, хотя закон ввел понятие «открытые данные».

Как облагать налогом платные сервисы, которые строятся на открытых данных?

Вернемся к тенденциям, нормам, привычкам и традициям пользователей. Опираясь на опыт Поднебесной, стоит явно отметить, что драйвером регулирования должен быть в первую очередь продукт с использованием данных, которые сегодня уже во многом представлены алгоритмами с использованием машинного обучения.

Наиболее популярные алгоритмы, которые стоит рассмотреть в регулировании:

• **Рекомендация и оптимизация контента** – такие платформы как Spotify, Netflix, Amazon, Ozon используют алгоритмы, которые постоянно наблюдают за нами. Они исследуют, как долго, часто и что именно мы смотрим, что нам нравится. Алгоритмы могут анализировать картинку, могут дать описание к ней, могут спроектировать картинку на основании наших любимых фильмов и актеров, чтобы продать нам новую услугу или товар.

• **Высокочастотный трейдинг** – такие специальные алгоритмы, которые перехватывают запросы пользователя, совершают покупку за тысячные доли секунды и продают их обратно пользователю, но с небольшой накруткой. Некоторые эксперты рынков капиталов называют HFT (высокочастотный трейдинг) паразитами, так как при выставлении ордеров пользователь фактически не может купить по текущей цене с рынка тот или иной инструмент. Рекомендую изучить историю ребят Investors Exchange[94] – они первые, кто придумал биржевую площадку, где нет таких пакостных посредников.

• **Реклама** – все поведенческие паттерны строятся на куках, ваших специальных идентификаторах внутри Интернета. Выбрали

книгу в электронном магазине о криптовалюте, передумали и ушли, потом зашли на видеохостинг, а в стоке вам показывают фильмы про криптовалюту. Совпадение? Не думаю. В Интернете как таковом нет своей памяти, но есть куки. Куки представляют собой ячейку памяти, только эту память видят все участники Интернета. Все рекламщики используют ваши куки, вы даже можете об этом не знать. Более того, они подсовывают свои собственные куки в ваши куки, чтобы сказать всем своим партнерам, что вы были у них. Существует много трекинговых методов, например флэш-куки, которые сложно удалить. Не всем хочется, чтобы кто-то знал, как в случае с Target, что вы ждете ребенка раньше, чем об этом узнают ваши близкие.

• **Поисковые запросы** – представьте, что выборы близко, а вы гипотетически никак не определились (опустим, что все уже предрешено). Попробуем перенестись в мир неограниченных возможностей и представить, что ваш голос еще по-прежнему учитывают. Вы вбиваете в поисковике имя потенциального **кандидата**[95], и вот уже сотни разных ссылок показывают информацию по этому человеку. И вроде бы все круто – теперь вы проинформированы, а значит, вооружены. Но представьте на минутку, что вам показывают один мусор, только плохие статьи и прочую грязь, так как именно в этот момент алгоритм поиска дал сбой. Как сильно это повлияет на ваше мнение о кандидате? Согласно одному из экспериментов в области психологии выбора, Роберт Эпштейн, психолог Американского Института Исследования и Технологий Поведения, выявил, что если человек в данной ситуации увидит позитивные новости о кандидате, то он проголосует за него с вероятностью более 48-ми процентов. И наоборот. Эпштейн назвал это явление Voting Manipulation Power – сила для манипулирования голосованием. Конечно, все это было в рамках лаборатории. Поэтому в 2014 году исследователи отправились на выборы в Индию, где в голосовании были задействованы восемьсот миллионов человек. Из них реально оказали участие порядка 430 миллионов. Исследователи провели очередной эксперимент на тех голосующих, которые не решились с выбором. В зависимости от того, какую информацию исследователи показывали людям, сила VMP оказывала на них влияние. Влиянию поддались от 24-х до 72-х процентов людей, участвующих в эксперименте. Впоследствии они назвали это эффектом Fox News – определяющим, как жители города, с консервативным кабельным каналом, становились более консервативными в своем выборе. В 2010 году Facebook мотивировал проголосовать более 340 тысяч человек, поставив на странице простую кнопку «I voted» (я проголосовал). Целевая

аудитория составила 61 миллион пользователей. Из них 611 тысяч (один процент) получило сообщение, стимулирующее к голосованию, в топ-списке новостей. Другие шестьдесят миллионов человек получили специальное «социальное» сообщение, практически аналогичное, за исключением того, что оно включало в себя профили проголосовавших друзей. Люди, увидевшие «социальное» сообщение, с вероятностью в два процента чаще кликали на кнопку «I voted», в отличие от первой группы, где реакция составила 0,4 процента. Такое «социальное» сообщение обернулось в шестьдесят **тысяч голосов**[96]. В среднем у пользователей Facebook в 2010 году было около 150-ти друзей, из них близкими были только десять (для территории США), и эксперты сошлись на мнении, что такие компании как цифровой «стук в дверь» могут увеличить количество избирателей на восемь процентов. Но что, если бы Facebook не ставил кнопку «I voted», а вместо этого бы просто попросил проголосовать за конкретного кандидата? Такой эксперимент предложил профессор **Гарварда Джоннатан Зитрейн**[97], назвав это феноменом цифровой **избирательной географией**[98]. Например, в январе 2012 года Google заменил главную картинку в поисковике на специальный логотип, при нажатии на который можно было попасть на страницу с подачей петиции против онлайн-пиратства. В результате этой петиции появился специальный законопроект (SOPA), который расширил полномочия американских правоохранительных органов и правообладателей и ввел жесткие наказания. Проект в итоге застрял в Конгрессе и так и не был принят, но факт остается фактом, траффик перенаправлялся и накачивал петицию.

Возвращаясь к вопросам регулирования, хочу в первую очередь отметить, что все существующее регулирование не рассматривает, к сожалению, пользователей с точки зрения их жизненного цикла, не учитывает продукты с использованием алгоритмов.

Большинство стран фокусируется на создании базовой инфраструктуры для управления рисками, нежели точного управления регулированием конкретных кейсов, в том числе и те, которые привел я, несмотря на то, что каждому из них уже более пяти лет.

Глава 5
Метаданные

И построил он замок.

Сижу смотрю фильм «**Анон**»[99], где общество лишено личной жизни и прав. Специальные службы записывают на сервера с терабайтами данных все, что происходит вокруг, используя наше зрение. Внезапно происходит череда загадочных убийств… Главный герой встречается со своим напарником, чтобы исследовать их, и вместе они начинают анализировать метаданные жертв.

В этот момент в моей голове не происходит сбоя, потому что я понимаю, что такое метаданные, и все в фильме выглядит крайне логичным. Но что, если я понятия не имею об этом? Тогда мне подсовывают красивый фантастический фильм по аналогии с «**Особым Мнением**»[100].

Глаза – это самый высокоскоростной интерфейс, поэтому они находятся на голове. Если бы они были, скажем, на заднице, то сигнал от них доходил был долго, и весь мир воспринимался бы нами с большой задержкой. Герои фильма «Анон» все делают глазами: звонят друг другу, оплачивают покупки, передают файлы того, что видят, идентифицируют себя. Все с помощью глаз.

Samsung представил в 2018 году, умную контактную линзу, которая может совершать эти операции с помощью **глаз**[101]. Линза проецирует изображение на глаз, позволяя воспринимать контент новым образом. Но также линза содержит и камеру, позволяя записывать видеопоток того, на что смотрит человек.

На протяжении всего фильма бравые парни пытаются отследить цепочку серверов и выследить метаданные.

И вот тут мое сознание дало сбой, потому что я не смог сформировать в голове единую платформу, на которой все живут. Возникла куча вопросов: как так вообще получилось, что всех на нее перевели.

По ряду причин я не сторонник централизованных платформ. Во-первых, делать их очень долго и дорого. Во-вторых, размер риска взлома гораздо выше, так как все находится в одном месте. В-третьих, чтобы управлять таким объемом данных, нужно правильно структурировать их, а это определенные компетенции.

В моем опыте был один проект построения крупного хранилища данных. Мы пытались совместить все ключевые функциональные подразделения организации вместе и получили

ситуацию, сложность которой не могли представить.

Представьте себе на минуту, что человек, который работает с большими объемами данных и проектирует сервисы, должен уметь разбираться в том, с чем именно он работает, – риски, продажи, бэк-офис, финансы и отчетность и так далее. По факту, таких людей единицы, поэтому централизованные системы обречены. В какой-то момент с ними никто не сможет разобраться.

В этом заключается интересный парадокс: чем больше люди хотят контролировать и чем больше они тратят ресурсов на централизацию, тем меньше в реальности они контролируют, и тем сложнее становится сама система. Выживут только небольшие управляемые компоненты.

Лазейки, оставляемые архитекторами таких платформ, похожи на небольшие тропинки, по которым идешь будто в потемках. Если мир полон красок, то эти лазейки переносят в пространство, где красок нет, но есть описание, что какие-то материалы применяются. Будто хоббит надел кольцо, и мир преобразился, лишился красок и стал похож, скорее, на чертежи.

Уж не знаю, какие еще аналогии привести, но смысл, думаю, понятен. Речь идет про те самые метаданные. Подложку мира. Описание того, как работает основная сцена.

Изучать эту подложку – это как смотреть на чертежи здания. Либо ты видишь, что архитектура безупречна, либо, что у архитектора руки растут из другого места.

А если таких зданий много, и вы находитесь в большом городе? Вдруг вы хотите открыть свой бизнес по продаже окон. Вам бы прикинуть, сколько окон вы можете продать и кому. Сможете просто взять и посчитать?

Возьмем что-то посложнее, например локомотив. Он состоит из секций, секции состоят из узлов, узлы представляют собой объединение деталей. Вот локомотив приехал на ремонт. Как понять, сколько конкретных болтов в нем нужно заменить в рамках регулярного ремонта? Нужно, чтобы техническая документация имела определенное описание, чтобы это описание можно было использовать и сделать запрос к информационной системе, где оно хранится. Бинго, правильно, используем метаданные.

Метаданные не только нужны для поиска и работы с большими массивами данных. Их еще очень часто используют различные люди и организации для получения доступа к тому, к чему они его легально получать не должны.

Большинство провайдеров сервисов (телекоммуникационные компании, мессенджеры и другие) собирают метаданные о

звонках и сообщениях. В случае с iMessage, такие сообщения будут содержать помимо времени звонка еще и данные о номере телефона, IP-адресе и номере адресата, который получил сообщение. Все это хранится в едином логе – истории изменения метаданных. Информация используется и предоставляется третьим лица, если на то есть решение правоохранительных органов.

В фильме «Анон» следователь получал доступ, используя метаданные к архивным файлам людей, и показывал их родственникам или пострадавшим жизнь глазами обвиняемых. И все это через метаданные, которые хранят последовательности с обращениями серверов о том, какие данные там хранятся.

Метаданные чаще всего используются для ведения каталога. Книжки, вещи, запчасти к велосипеду. В общем, если запускать большой бизнес, который будет торговать или что-то предлагать в Интернете, то работа с метаданными – это из ряда must have.

Откуда она такая появилась

В 1965 году одним из первых появился стандарт **IPTC-описания фотографий**[102] в Ассоциации Новостных Газет Америки. Стандарт включал в файл обязательную информацию – автора, заголовок, дату создания. Когда файл обрабатывали приложения, появившиеся чуть позднее (к примеру, Photoshop), они уже адаптировали этот стандарт, и все изменения сохраняли в файлики с форматом *JPEG, *PNG или *TIFF.

Позднее, в начале 90-х, все это было уже структурировано более изящно и преобразовано в XML, чтобы можно было работать со сложным поиском. Adobe показал миру платформу XMP, которая встраивала мета-информацию в файлы самостоятельно, без участия пользователя в формате, который мог быть воспринят как человеком, так и машиной.

Метаданные стали ключом к поиску. К сложному поиску в больших массивах данных.

Самое важное здесь, конечно же то, что формат данных по-прежнему понятен и человеку, и машине.

Стандарт XMP впоследствии стал открытым стандартом ISO (16684-1). В совокупности такая стандартизация дала возможность:

• Управлять контекстом не только во время путешествия вокруг баз данных, но и обмениваться их в индустрии в целом.

• Организовать полноценный поиск среди разных форматов файлов и различного контекста.

• Управлять и определять взаимосвязями объектов, их жизненного цикла (создание контента и его удаление).

Почему я начал с фото, а не с книг, где метаданные

появились в первую очередь? Например, с той же **Десятичной Классификации Дьюи**[103],которая впоследствии стала основным иерархическим справочником материала для библиотек на добрые полтора века.

Просто потому, что фотографии сыграли ключевую роль в обучении алгоритмов данных. Большая часть нейронный сетей обучалась изначально на большом массиве фотографий, где метаданные были размечены вручную – Image.Net. Собственно, если бы не эта большая работа, алгоритмы не могли бы отличать красное от черного, собак от кошек, человека от унитазного бачка…

Сегодня метаданные к картинкам теперь могут заполнять сами алгоритмы. То есть они прямо анализируют картинку. Пусть, скажем, там изображена собака на прогулке вдоль городской улицы. Алгоритм напишет, что на фотографии – собака, порода – лабрадор, а также там есть дома, пожарный гидрант, и все происходит днем. Теперь это тоже часть метаданных.

Само слово «метаданные» впервые ввел Филипп Бэглей в своей книге «Extension of Programming Language Concepts», опубликованной в 1968 году. Он и ввел понятие «прескрипторы», которые описывали данные кратким и понятным образом.

Теперь такие данные собираются обо всем, начиная с географических карт, заканчивая музыкальными файлами.

В России стал популярен сериал «Карточный домик», он о том, какие нелегкие дела творятся в Белом Доме США. Сериал встал наравне с таким гигантом, как «Игра престолов». Пока я работал в офисах последние несколько лет, за обедом коллеги обсуждали исключительно то, как разворачиваются дела у Фрэнка Андервуда, одного из главных героев политического триллера. Но кто задумывался о том, что Netflix инвестировал в создание этого сериала анализ метаданных от 44 миллионов своих **пользователей**[104]?

В общем, понятно, что метаданные – это важно, и что на них делают бизнес.

Правда, сегодня существует ряд проблем, связанных с ними. Вот ребята делают исследование. Их больше всего интересует его результат, нежели сделать по итогам исследования правильную архивацию, расставить необходимые метки и **признаки**[105]. Получается, что исследование есть, а поженить это исследование с другими наборами данных – сложновато.

Порой даже те данные, которые публикуют, представляют собой не полную выборку, а какой-то ее определенный фрагмент для поддержания результатов исследования.

В 2016 году журнал «Science» опросил более полутора тысяч ученых, представляющих ключевые дисциплины (химия, биология, медицина, физика, экология и другое), с целью понять, как именно они используют свои данные: могут ли они их воспроизвести и повторить результат исследования?

Оказалось, что более семидесяти процентов исследователей не смогли воспроизвести результаты других ученых. Как факт **52 процента**[106] из них подтвердило, что в науке начался новый кризис воспроизводимости результатов, говорящий о том, что большинство результатов полученных в современной науке невозможно повторить. Одна из названных причин – данные, другая – отчетность, публикуемая в поддержку тех или иных гипотез. Такая отчетность представлена выборочно, то есть команда проекта публикует только те отчеты, которые поддерживают исследование, а не опровергают его.

Низкий уровень культуры работы с ними делает невозможным повторное их использование. С другой стороны, в науке нет консенсуса в вопросе «как нужно использовать данные, чтобы можно было возобновить на них результаты исследований другого учетного». Просто потому, что время, потраченное на причесывание таких данных, увеличивает время, потраченное на исследовательский проект, более чем на тридцать процентов, поэтому не всем очевидно, зачем это делать.

Одна из успешных стратегий снижения риска заключается в том, что на основании данных нельзя будет ничего воспроизвести – в этом случае необходимо формулировать гипотезу и планировать, какие данные нужны будут для ее подтверждения с участием третьих лиц.

Какова разница между воспроизведением и репликацией результатов исследования? Для репликации можно взять данные из репозитория и использовать на них код. Пусть это не всегда работает, но все же работает. А вот получить результаты, близкие к лабораторным, крайне сложно, потому что нет единой модели метаданных – то есть, описания того самого единого города, в котором мы находится.

Представьте, что у вас есть чертежи различных строений на разных языках с разным форматом описания. А вам нужно попробовать выделить общее между ними, например, понять, где находится лифт, нуждается ли он в ремонте, узнать, как организованы системы снабжения и коммуникации. У вас на столе лежит несколько различных схем, в которых без бутылки не разобраться.

А что, если часть таких схем просто сфабрикована и не имеет ничего общего с реальными построениями?

Йошихиро Сато был известным уважаемым японским специалистом по костям. Он посвятил пятнадцать лет исследованиям в области остеохондроза, опубликовал порядка двухсот научных результатов и провел более 33-х клинических исследований.

В составе группы ученых Йошихиро Сато исследовал влияние болезни Паркинсона на снижение массы костей и скелета, как следствие, на возможное осложнение **остеохондроза**[107] у пациентов в районе Kahanzan. Формировались небольшие выборки пациентов по 86 человек в среднем, к которым добавляли других пациентов с болезнью Паркинсона, и давали лекарство вместе с таблеткой плацебо. Средний возраст пациентов составил 70,6 лет (от 65 до 88). В контрольной группе было 35 мужчин и 51 женщина.

Группы наблюдались в течение 18-ти месяцев, пациенты оценивались и осматривались каждые две недели. Им давали витамин D, а сложное рентгеновское оборудование анализировало толщину их костей. Во избежание влияния третьих факторов составлялись сложные опросники по диете, влиянию солнечного света и так далее. У одной из групп было выявлено существенное снижение кальция в костях (более чем на 25 процентов). Тогда этим пациентам предложили использовать определенные витамины и питание для повышения усвояемости витаминов в костях, чтобы не было потери массы. Исследования показали положительную корреляцию употребления витаминов D и B для пожилых людей со сложными болезнями Паркинсона и Альцгеймера.

В январе 2017 года Йошихиро Сато скончался при невыясненных **обстоятельствах**[108]. Оказалось, большинство его работ по клиническим исследования содержат сфабрикованные данные, которые впоследствии были опровергнуты научным сообществом. Он стал автором крупнейшего скандала в науке с подделкой данных.

Годом ранее Марк Болланд из университета Окланда (Новая Зеландия) провел статистические исследования с использованием данных господина Сато за все 15 лет работы и выявил, что большинство его исследований – подделка. Даже соавторы в большинстве работ, как оказалось, не знали о своем участии и не участвовали в этих работах вовсе. Теперь ответ, каким образом Йошихиро Сато смог опубликовать более двухсот научных работ, лежал на поверхности. При более детальном изучении данных, открывались новые подробности о том, как он мог собирать 280 пациентов для своих исследований всего за два месяца или наблюдать 780 пациентов в течение 18 месяцев одновременно.

Для сравнения отмечу: нанять на работу 280 сотрудников за два месяца возможно, разве что для простой работы, например, контакт-центр или поддержки. Найти квалифицированных специалистов или, как было указано в исследовании, пациентов с конкретной болезнью – очень проблематично.

Судьбы людей вершились исключительно при использовании данных. Болланд никогда лично не встречался с Сато и впервые о нем услышал только в 2012 году, когда его коллега доктор Алисия Авенелли рассказала ему о странных данных в исследованиях **Сато**[109], которые при проверке оказались слишком научными.

Первый контакт состоялся в Марте 2013 года, когда Болланд и Авенелли написали в журнал Американской Медицинской Ассоциации – наиболее уважаемый журнал из тех, которые публиковали статью и исследования Сато. Главный редактор журнала дал указание обратиться к Сато и его институту, чтобы получить пояснения по выявленным в данных фактам.

Через два года, в апреле 2015-го, никакого ответа не пришло, в связи с чем журнал опубликовал результаты расследования и претензию к полученным и ранее опубликованным результатам Сато. Репутация Сато была настолько высокой, что журналы не решались поначалу идти против него и предъявлять обвинения в искажении результатов.

К декабрю 2016 года только 10 из 33 опытов были опровергнуты, когда вышло очередное расследование в журнале «Нейрология».

Только пять **процентов**[110] из опубликованных исследований приходят из Японии, поэтому такой удар по научной среде привел к потере репутации для японских ученых. Остается загадкой, зачем Йошихиро Сато подделывал так много результатов своих работ и фальсифицировал данные.

На сегодняшний день он занимает шестую строчку по количеству отозванных результатов клинических **исследований**[111].

На первом месте в этом списке находится японский ученый Йошитака Фуджи, который занимался клиническими исследованиями в области анестезии. Согласно отчету, опубликованному 8 марта 2012 года, во всех 169-ти клинических испытаниях данные были искажены и сфабрикованы (в общей сложности для 171-го исследования).

На втором месте находится Хоаким Болд с исследованиями в области грудной хирургии, который так же был уличен в подделке **данных**[112].

На третьем месте – Дидерик Штапель со своими

сфабрикованными исследованиями в области социальной **психологии**[113]. В целом масштаб таких проблем в науке поражает. Эти имена – лишь верхушка айсберга.

Спасти эту ситуацию может блокчейн. Одно из решений – платформа **Frankl**[114], которая интегрирует всех ученых в единую открытую сеть. Туда можно загружать данные и делиться ими друг с другом для проверки чужих или проведения своих подобных исследований. Если не вдаваться в подробности, то Frankl пытается создать распределенную сеть, где можно будет контролировать качество данных, что фактически снизит размер потенциальных фальсификаций.

Регистрировать все метаданные на блокчейне – самый простой шаг, но очень мощный, чтобы контролировать полноту данных, используемых в исследованиях.

Итак, метаданные – это в первую очередь явление чисто человеческое, то есть, его нет в природе. Человек разработал его специально для себя, чтобы обрабатывать большие объемы информации и оптимизировать поиск необходимого контента. Метаданные уже спроектированы и во многом генерируются автоматическими устройствами.

С другой стороны, мы вовсе не коснулись проектирования баз данных. И это хорошо, потому что это очень занудная для обычного читателя тема. Если кратко, то при проектировании сложных экосистем метаданные используются для управления потоками загрузки и обработки данных. Они формируют управляющую логику того, как данные собираются и обрабатываются.

Есть интересная работа, надеюсь, не поддельная, по оптимизации работы с базой данных **Википедии**[115]. В работе предложен специальный инструмент по управлению и архивированию исторических данных: индексы, каталоги, описание – все, что помогает оптимизировать поиск по историческим данным.

В зависимости от используемого решения систем хранения и обработки данных, на рынке предлагаются различные решения по управлению метаданными, использующими специальные **сервера**[116]. По версии «волшебного квадранта» Гартнера, лидером таких решений является **Informatica**[117]. Хотя, конечно, я слышал, что за то, чтобы попадать регулярно в этот квадрант, нужно платить определенную сумму, поэтому там нет начинающих или малоизвестных компаний.

Все эти решения отличаются как функциональными возможностями, так и пользовательским интерфейсом. Пользователями таких решений являются инженеры в области

данных, они здесь самый ценный ресурс, так как этой компетенции, к сожалению, не обучают в ВУЗах, а количество специалистов на рынке стремится к минимуму.

Раньше процесс найма проходил в основном самостоятельно, в недрах IT. Сегодня за это должен отвечать отдельный лидер в организации. Но вопрос о том, где взять специалистов, по-прежнему актуален, поэтому приходится выкручиваться. Я, например, был сторонником того, чтобы поощрять горизонтальное движение сотрудников как внутри организации, так и за ее пределами.

Мы собирали ребят из службы IT-поддержки, потому что им по факту приходилось ковыряться в базах данных различных IT-систем, анализируя те или иные метаданные. Приглашали на работу сотрудников других компаний, которые занимались выпуском и проверкой финансовой отчетности. Такие люди понимают ценность данных и анализируют, в каких системах лежат наиболее ценные данные. Каждый такой кейс мы рассматривали отдельно.

Обучение новым навыкам мы строили на основе практики, потому других источников знаний у нас не было. С одной стороны, это создавало риски, с другой – поощряло свободу к действиям. Сотрудники были как никогда нацелены на результат, а их предыдущий опыт помогал находить нестандартные решения в тех или иных вопросах.

Стоит отметить, что бизнес-лидеры не всегда понимают ценность отдельно взятых решений по работе с метаданными.

Это какая-то малопонятная область работы и применения ресурсов, и не всегда ясно, зачем на это нужно тратить время. Надеюсь, что пример с чертежами зданий и новые фильмы по аналогии с «Аноном» позволяют раскрыть потенциал метаданных. Моделей монетизации таких решений очень мало.

Например, при расчете себестоимости функции работы с данными как сервиса, я использовал исследования Калифорнийского университета, где была приведена модель затрат и ценообразований функции использования данных. В этом отношении я мыслил достаточно просто – нужно было продавать именно данные как сервис, а работу с метаданными сделать обязательным компонентом себестоимости этого сервиса. Сервисная модель работы с данными – относительно новое явления для бизнеса, так как большинство лидеров для тех или иных задач выделяют ресурсы напрямую.

Сервисная модель предоставления данных – это отражения новой идеологии, постепенной захватывающей новые ниши в различных секторах экономики. Эта идеология называется

концепцией единого цикла, – конечному клиенту предоставляется не сам продукт, а результаты работы этого продукта как сервис. Можно не покупать автомобиль, а платить за эффективный километр. Так и здесь. Нет смысла продавать сложную инфраструктуру, нужно предложить сервис доступа и получения данных.

Сервисы работы с метаданными можно также выделить в отдельный сервис для поддержки работы и обучения нейронных сетей. С другой стороны, с использованием метаданных можно сделать отличный сервис по получению и сбору данных из различных источников, так называемые **краулеры**[118] для создания хабов данных. Такие краулеры помогают собирать различные данные из сложно структурированных источников, таких как сайты, сложные файлы, внешние хранилища и другие.

Из метаданных выстраиваются те самые связи, благодаря которым герои фильма «Анон» могли перемещаться между различными источниками данных. Вспоминая этот фильм, я в первую очередь представляю объем работы, которую проделало человечество, чтобы связать источники данных, сделать сложно иерархические структуры, эффективно применяющие алгоритмы быстрого поиска.

Но, с другой стороны, именно метаданные становятся тем самым уязвимым местом в системе, если за ним никто не присматривает. Хакеры могут использовать эти данные для получения незаконного доступа, а проблемы с качеством могут поставить крест на важнейших исследованиях для человечества.

Все так красиво и понятно. Строить здания без чертежей – как бы неправильно, но что же делать со всей существующей инфраструктурой, где место для метаданных могло быть не заложено вовремя в виду тех самых коммуникационных сложностей?

Необходимо искать нестандартные пути и решения, потому что эффективное управление информационным ландшафтом – это ключевой вызов сегодняшнего дня, на который крупному бизнесу еще предстоит ответить.

Глава 6
Зачем нужно качество данных?

Раз вы добрались до этой главы, тяга к новым знаниям дарована вам природой или же воспитана в суровых сибирских условиях.

Тема качества данных лично меня коснулась не сразу. Работая в команде вышколенных аудиторов – выпускников лучших ВУЗов страны, мы привыкли опираться на свое мнение: делать выводы на основании данных, документировать и предоставлять описание своих ключевых суждений, которые ложились в основу аудиторского заключения.

Вот ты приходишь в банк, и тебе дают на руки выгрузку из проводок оборотно-сальдовой ведомости – это такая большая табличка с остатками и оборотами, где находится все на свете. И обычно аудиторы сбивают сначала оборот и остаток по оборотно-сальдовой ведомости с тем, что есть на счетах в отчетности, которую банки публикуют и отправляют в Центральный банк Российской Федерации.

Итак, вот мне сгружают тонну данных – это примерно несколько миллионов записей. В то время Excel еще не умел работать с такими количествами, это уже после он смог анализировать четыре миллиона записей. Приходилось разбирать этот большой объем данных на части.

Каждому аудитору доверяли какую-то конкретную секцию. Если смотреть на финансовую отчетность, то секция – это один ее раздел. Надеюсь, вам удавалось хоть раз на нее взглянуть. Если нет, взгляните для примера на публикуемую отчетность по МСФО[119] любой российской компании или банка, например ВТБ, она, кстати, отличается от остальных тем, что ее составляют в миллиардах рублей. ВТБ был одним из первых банков, которые перешли на выпуск отчетности в миллиардах рублей. Сегодня не так много компаний могут этим похвастаться.

Что это значит в практическом плане? В первую очередь это, конечно же, размер потенциальной ошибки. Раньше у аудиторов была проблема с одной частой ошибкой при подготовке отчетности – с округлением.

Вот представьте, что вы округляете цифры для отчетности так, чтобы это соответствовало принятому размеру внутри самой отчетности – миллиарды, миллионы, тысячи и так далее. Раньше, когда появлялась та самая «единица», которая возникала из-за округления значений строк, аудиторы обычно пристраивали ее в

какую-то из строчек так, чтобы общая сумма сходилась. Потому что иначе при сложении сумма расходилась с той суммой, которая была посчитана на более маленьких значениях.

Да и какая разница, в какой строчке баланса будет больше на одну единицу, а в какой меньше. А если дело касается миллиардов? У вас из-за округления появится плавающий миллиард…

Насколько сильно это повлияет на качество конечных данных? Насколько сильно это повлияет на принимаемые решения?

В таком случае аудитор всегда обязан определить ту разницу, которая может и не может одновременно влиять на принимаемые решения на основании выводов из финансовой отчетности. Вы спросите меня как?

Очень просто, есть такое понятие как «материальность» или по-нашему – существенность. Это означает размер потенциального искажения информации, которая может ввести пользователя в заблуждение, из-за которого он сделает неверные выводы, а то и неверные действия на основании некорректных данных.

Насколько материален миллиард? Вот вы смотрите на отчетность, возможно, вы ничего в этом не понимаете, но вам важно, что тут «плавает» миллиард между строк?

А теперь давайте спустимся на уровень ниже, так как это все вершина пирамиды. На самом нижнем уровне данные собираются из разных источников и могут искажаться гораздо существеннее чем на один миллиард рублей.

Но стоп…

Вы меня спросите, как такое возможно, если отчетность не сходится всего на один миллиард рублей? Следите за руками. Когда аудитор делает проверку, он должен гарантировать, что цифры не искажены во всех материальных аспектах. Материальность можно определить по-разному. Например, взять один процент от размера полученной прибыли до налогов или полпроцента от размера активов, которыми владеет организация и так далее. В общем, подходов много. Но суть в том, что, определяя этот самый размер «существенной» чувствительности к принимаемым решениям, организация отсекает ниже этого уровня все суммы, которые отныне считаются неважными[120]. Таким образом аудитор проверяет отчетность во всех материальных аспектах и гарантирует при этом достаточный уровень уверенности в положительном исходе. Если спросить меня, что такое «достаточный» уровень, и чем он отличается от «абсолютного», то я отвечу – всем. Это не одно и то же.

Иными словами, аудитор не может проверить все цифры и все данные в организации. Все, стоп. Зачем я это описываю?

На самом деле, я так попытался пояснить, что такое искажения данных и подготовить вас к сложной части. Но теперь давайте прыгнем в эту темную пучину и попробуем разобраться. Хочу, чтобы вы понимали, что любая ошибка в данных – это финансовый эффект, независимо от того, сделана она на верхнем уровне подготовки самой отчетности, или же эта проблема была в первичных данных.

Итак, двигаемся дальше.

У меня есть брат. Мы близнецы. И, наверное, как стало понятно, мы однофамильцы. Так случилось, что мы имели счет в одном и том же банке. Не буду говорить в каком, этому бренду и так досталось, но особо пытливые поймут.

В один прекрасный день в этом самом банке меня внезапно «склеили» с другим клиентом так, что, открыв приложение этого мобильного банка[121], я увидел остатки на счетах другого человека, и у меня даже была возможность управлять этими счетами. Хорошо, что это были средства моего брата.

Но погодите-ка. Причем тут он? Вот и мне это интересно. Но система в банках внезапно заменила все мои данные на его. Даже при звонке в контакт-центр с моего телефона у них отражался другой человек, совсем не Алексей. Опустим, сколько проблем пришлось решить, чтобы я снова мог видеть свои счета и пользоваться банковскими сервисами.

Такая проблема бывает не только у меня, и она никак не связана с тем, есть ли у вас брат-близнец. Она даже может быть не связана с конкретным банком. Оказывается, это ошибка в данных, которая вызывает коллизии во внутренних IT-сервисах, делая недоступным для конечного потребителя стандартный набор функций, на которые он рассчитывал.

Тема качества клиентских данных всегда особо актуальна. Однажды у меня была проблема с одним из клиентов, который по одной из систем имел категорию «хороший клиент», а по другой – был «террористом». «Террорист» означает, что система нашла совпадение с одним человеком из списков, которые публиковал Центральный банк. Не все участники этого списка обязательно террористы в буквальном смысле, они вполне могли просто нарушить какой-то закон и попасть туда. Какое-то время назад, как мне рассказывали, в такие списки попадали люди, нарушившие законы, запрещающие участие в пикете или демонстрации.

Что банк должен делать в случае, если один из его клиентов попал в такой список? Правильный ответ: в соответствии с

действующим законом он должен приостановить банковское обслуживание и закрыть банковский счет. Кстати, именно эту отговорку часто используют банки, если им нужно закрыть счет и особо не навлекать на себя сложности и тяжелые разбирательства.

А что же должен был делать я? Закрывать счет или нет?

Как бы вы поступили?

Разбирая проблемы в клиентских данных, можно отметить известный случай 2015 года, когда Федеральная налоговая служба ввела для всех налоговых агентов[122] новое обязательное поле к заполнению при подаче справки 2-НДФЛ[123]. Этим полем стало поле «ИНН», которое каждый налоговый агент обязан был заполнять во избежание штрафа.

Причем тут банк? Все просто. Вот я взял кредит в банке и больше не могу по нему платить. Что делает банк? Помимо того, что он насылает на меня своих демонов-коллекторов и бегает с требованиями погасить задолженность, после определенного периода он вынужден будет эту задолженность списать.

Но самое интересно будет дальше. С точки зрения буквы закона и учета, списание задолженности означает, что эти кредитные деньги я признаю себе как доход, а значит, я должен с них заплатить налог. Но если денег у меня нет, то, естественно, налог платить должен банк.

Маленький дурдом получается. Мало того, что банк терпит фиаско с клиентом, так он еще и вынужден заплатить 13 % с этой сделки в бюджет. C'est la vie. Не будем разбираться в справедливости сего факта, попробуем разобраться в следующем. Вот банк должен уплатить налог, а значит, подать еще и справку 2-НДФЛ в ФНС, а в этой справке теперь стоит обязательное поле «ИНН».

Но вот незадача, при выдаче кредита, банк не узнал эту информацию у клиента, потому что она не являлась обязательной. Существует много различных случаев получения такого дохода.

Доход по депозиту тоже можно признать таким же доходом, по которому обязаны подать справку 2-НДФЛ, если ставка, по которому он начисляется, в определенное количество раз больше, чем ставка рефинансирования[124], то в этом случае банк должен отчитаться также в ФНС. Получается, что если вы видите высокую ставку по депозиту и думаете туда положить денежку, то знайте – это сверхдоход, по нему нужно удержать налог. Удерживание налога – это обязанность налогового агента, то есть банка. Но разве кто-нибудь спрашивает у вас поле «ИНН» в момент, когда на каком-нибудь банковском сайте вы размещаете свою заявку на

депозит?

Конечно, нет. Такое поле по-прежнему не является обязательным при открытии договора депозита или вклада в том или ином банке.

В 2015 году ФНС грозился, что штраф за незаполненное поле «ИНН» с каждой такой записи составит двести рублей, а с 2016 года вырастет до пятисот. Представьте, что я банк. И если у меня сто тысяч клиентов, по которым я списал задолженность, или по которым начисляют зарплату с моих банковских карт, то, умножив сто тысяч на двести рублей, я получу ежегодный штраф в размере двадцати миллионов рублей.

Оценка грубая, но, тем не менее, думаю, смысл стал понятен. С какого-то момента плохое качество данных организации, с которым она работает, начало создать проблемы и новые штрафы для самой организации. Иными словами, плохие данные стали «токсичным» активом, приносящим организации новые убытки в будущем.

Сколько таких «важных» полей существует внутри IT-систем большой организации? Отвечу досрочно – десятки тысяч как минимум. Десятки тысяч таких полей, десятки тысяч гигантских табличек, которые нужно проверять и контролировать.

Ошибки могут быть разными. Иногда важно, чтобы указанные данные существовали в реальности, как в случае с полем «адрес», чтобы банк смог доставить корреспонденцию своем клиенту. Многие вбивают в это поле все, что вздумается, но хорошо, что есть такие отличные сервисы как DaData, которые не позволяют вбить несуществующий адрес. Об этом впереди.

Наверняка, заполняя какую-нибудь формочку на сайте, вы сталкивались с его просьбой указать «индекс». А потом вы нервно начинали гуглить индекс указанного адреса.

Но фишка в том, что «индекс» как поле можно и не запоминать, это атавизм. Из утвержденных публичных справочников типа ФИАС или **КЛАДР**[125], в правильном существующем адресе уже есть индекс, и его можно взять оттуда.

Это ведь просто прекрасно – не заполнять поле «индекс». Так почему же его до сих пор заполняют и спрашивают?

Оказалось, что в базе **ФИАС**[126] (государственный источник) поле «индекс» заполнено не совсем корректными почтовыми индексами. Нужно искать еще и другие правильные источники, например, базу данных «Почты России», но даже в этой базе нет всех тех индексов, которые есть в ФИАС.

Чтобы в этом всем копаться, нужно все это любить и получать удовольствие от разгребания подобных проблем. Большинство людей, ежедневно сталкивающихся с теми или иными цифровыми

сервисами, не знают о том, какой объем работы проводится для упорядочивания данных перед тем, как показать их клиенту.

Ошибки, опечатки и погрешности влияют на многие факторы в организации. Однажды, объединяя клиентов одного банка с другим, в процессе консолидации мы внезапно выявили, что несколько тысяч мужчин внезапно изменили пол и стали женщинами. Конечно, у нас демократичная страна, но сей инцидент произошел в суровых сибирских районах, поэтому я по-прежнему склонен думать о наличии очередной ошибки в данных.

Как быть? Как исправить ошибки, которые уже случились? Я ведь не могу пойти на «Горбушку» и купить компакт-диск с данными[127].

Для обогащения данных клиентов и заполнения поля «ИНН», мы пробовали различные методы. Звонили и спрашивали, просили прийти в офис и заполнить анкету, делали даже такую доработку в мобильном приложении или интернет-банке[128].

Эффект на общем потоке составил какой-то мизерный процент, то есть люди не шли и не давали свои данные. Пришлось менять банковские процессы и делать поле «ИНН» обязательным для всех продуктов. Такая вот головная боль из-за одного поля, а таких полей, повторюсь, очень много.

Кстати, если вдруг вы торгуете ценными бумагами, мало ли меня читают такие умные люди, то, наверное, обратили внимание, что в личном кабинете брокера, через который вы торгуете, появилось обязательное требование заполнить поле «ИНН». Совпадение? Не думаю.

Основные методы управления качеством данных

Качеством и проблемами в данных должны управлять специальные люди в организации.

Да ладно вам, согласитесь, что это уже очевидно. Будет странно, если я начну распинаться и объяснять слишком очевидные вещи.

Лучше я объясню, где и как организовать работу таких людей.

Для начала нужно все разбить на две части.

Первая часть – это так называемая служба поддержки пользователей, которая приходит на помощь, если что-то случилось с их данными. По-умному эта команда называется дата-стюарды. Да, именно так и называются – дата-стюарды. Русского перевода нет. Пусть будет это новое английское слово в нашем лексиконе. Привыкайте, дальше будет веселее.

Задача дата-стюардов проста и понятна – разобраться оперативно в каше под названием данные так, чтобы ни один клиент не пострадал, или чтобы вовремя выпустилась отчетность. Ну, кажется, понятно объяснил. Если нет, то, скорее всего, во время чтения этой книги вы слышали какие-то посторонние звуки или встретили другие препятствия, – настоятельно рекомендую избавиться от них. Главное, не говорите потом, мол, Алексей, ты пишешь «непонятно». Даже не думайте. Я ведь очень стараюсь.

Дата-стюарды работают по понятной и прозрачной методике: на них сыплется поток плохих данных, в котором им нужно быстро ориентироваться. В идеале они должны выполнять простые операции, поэтому для их эффективной работы нужно составить простую и прозрачную методику, где будет указано, что они должны делать в конкретном случае. Все. Баста.

Я обнаружил, что самая эффективная реализация функции дата-стюардов находится в бизнес-подразделении[129]. Дам вам десять баллов, если сможете засунуть их в департамент разбора жалоб клиентов, где они на месте смогут разбираться в проблемах с качеством клиентских данных.

Да, именно «там», где пишут жалобы через сайтик, контакт-центр или от руки (на бумажечке), должны сидеть дата-стюарды, которые работают с клиентскими данными.

К сожалению, все проблемы дата-стюарды решить не могут, иначе бы они рассыпались на много маленьких некрасивых кусочков.

На сцену выходит вторая боевая группа, которую я называю дата-инженеры. Как, надеюсь, понятно из слова «инженер», эти ребята должны что-то строить и проектировать. И все действительно так: они проектируют единую архитектуру управления качеством данных (проверки, средства контроля и, конечно, дизайн самих IT-систем и цифровых интерфейсов[130]). Если сравнивать обе команды, то инженеров должно быть меньше, потому что это более высококвалифицированные единицы.

Точка дислокации этой группы может быть разной. В организациях, где директор по данным (CDO) не является ключевым сотрудником компании и находится где-нибудь на уровне Board-2[131], такая команда может находиться как внутри финансового блока, так и внутри IT-блока.

Эффективнее будет расположить инженеров ближе к команде IT-блока, так как эти ребята должны непосредственно изучать IT-системы, углубляться в процессы и предлагать решения.

В случае, когда организация слишком продвинутая и зрелая, CDO обычно выступает уже на уровне Board – входит в Правление или стоит максимум на одну ступень ниже, но по-прежнему остается ключевым сотрудником.

Теперь вопрос. Где взять таких людей для позиции инженера по качеству данных на рынке? Если начать поиски на HeadHunter или на других подобных порталах, то достаточно быстро станет ясно, что количество релевантных кандидатов для обеих позиций оставляет желать лучшего.

Я выкручивался тем, что брал людей из других подразделений, создавая таким образом возможности для горизонтального роста. Происходило это как внутри компании, так и с привлечением соискателей из вне.

Инженеров я искал и брал как из службы IT-поддержки, так и из смежных департаментов, где есть бизнес-экспертиза. В службах IT-поддержки я руководствовался тем, что люди при разборе тех или иных инцидентов и проблем в IT-сервисах, вынуждены погружаться в то, как работают сами IT-решения.

Бизнес-экспертиза необходима, потому что инженер должен понимать, как влияет на бизнес выявленная ошибка в данных.

В качестве инструмента сведения и демонстрации общего эффекта проблем с качеством данных, я выбрал подход так называемого документа «аппетит к риску».

«Аппетит к риску» – это специальный документ, отражающий размер потенциальных рисков, которые несет в текущий момент организация, он политика относительно текущей позиции и ее фиксирования.

История документа лежит в Базельских требованиях. Это требования, которые были приняты международным сообществом (Базельским комитетом) как достаточные требования для управления банковским капиталом, чтобы иметь возможность управлять риском банковского банкротства (дефолта), а точнее, иметь возможность его предсказывать.

Так вот, «аппетит к риску» – это внутренний документ, который готовит организация для того, чтобы в момент измерить, достаточно ли у нее капитала для покрытия возможных убытков, которые могут возникнуть, если что-то пойдет не так. Сюда в том числе относятся действия и самой организации, когда она решается на более рискованные стратегии или запуск сложных продуктов. Надеюсь, кратко, но понятно. Этот документ обычно режет глаз, он должен показывать то, что у организации больше нет возможностей для маневров, и служить отрезвляющим душем для менеджеров.

После участия в проектах управления Базельскими требованиями внедрения сложных механизмов расчета достаточности капитала, мне показалось интересным применить логику построения этого документа к оценке рисков, которые содержат в себе некачественные данные.

И если раньше, когда я приходил в Правление организации, мне было сложно объяснить важность управления качеством данных, то, создав такой документ, где мы наглядно отразили размер потенциальных убытков от плохих данных (те же штрафы за поле «ИНН»), я смог сфокусировать внимание менеджмента на самом важном. На бабках.

Только такой язык сегодня понятен внутри бизнеса. Если ты явно покажешь, что качество данных создает убыток для организации, с высокой степенью вероятности, ты получишь новый ресурс для решения этой проблемы или митигации этого риска.

Есть много подходов и визуализаций отражения и создания классического документа «аппетит к риску». Все они применимы к созданию аналогичного документа в отношении качества данных.

В классическом варианте такой документ выделяет виды капитала (регуляторный, экономический) и виды рисков, но что же он должен показывать в отношении данных?

Как измерять качество данных?

Не буду претендовать на уникальность и вернусь к работе аудиторов. Уж очень она яркий отпечаток оставила в моей памяти. Мое вечное желание все вокруг связывать и комбинировать в поисках прорывных решений не дает мне об этом просто так забыть.

Как проверить что конкретная цифра в весьма конкретной отчетности верна?

Оказалось, очень просто: аудитор в своей работе использует так называемые «assertions» или «допущения», которые разбиты на определенные группы, коих опять конечное количество. Есть такой стандарт по международному аудиту **номер 315**[132], которым обязаны руководствоваться международные компании по аудиту. Так вот он говорит, что этих самых «assertions» в части финансовой отчетности всего 13 штук и они все поделены на три определенные группы.

Первая группа таких допущений относится к транзакциям и формируемой прибыли:

1. Наличие (Occurrence) – транзакция или событие действительно имело место и реальности произошло.

2. Полнота (Completeness) – транзакции, которые произошли, были отражены полностью.

3. Точность (Accuracy) – все данные касательно транзакций отражены без искажений.

4. Срез (Cutoff) – транзакции произошли в правильном отчетном периоде.

5. Классификация (Classification) – транзакции были отражены на правильном счете и правильной строчке.

Вторая группа уже касается остатков и самого баланса, и выглядит она следующим образом:

1. Существование (Existence) – актив, обязательство или указанный капитал действительно существуют.

2. Права и обязанности (Rights and Obligations) – то, что отражено в отчетности, и организация непосредственно это контролирует.

3. Полнота (Completeness) – все, что реально существует, все это отражено полностью во всех соответствующих строчках отчетности (активы, обязательства, капитал).

4. Оценка и распределение (Valutation and allocation) – все, что реально было, отражено корректно с точки зрения оценки этих объектов. К примеру, ценные бумаги, которыми владеет организация, должны быть отражены по самой последней

рыночной котировке и так далее.

Третья группа уже касается непосредственно раскрытий и пояснений финансовой отчетности:

1. Существование (Occurrence) – все, что было раскрыто и пояснено в отчетности, оно действительно случилось. Если в отчетности написано, что сгорел завод, значит, он действительно сгорел.

2. Полнота (Completeness) – все, что в реальности было, тоже раскрыто. Если еще сгорел амбар помимо завода, и это важно, то это нужно раскрыть.

3. Классификация и понимание (Classification and understanability) – вся финансовая информация должна быть представлена таким образом, чтобы было все просто и понятно. Никаких сложных раскрытий и сложных описаний.

4. Точность и оценка (Accuracy and valuation) – все посчитано честно и аккуратно.

Надеюсь, сознание после этих трех групп еще не потухло, идем дальше. Когда работает аудитор, он оценивает материальные искажения в каждом из этих тринадцати измерений. Материальность отклонения также определяют по разным правилам. Так управляется риск контроля качества информации в финансовой отчетности.

Мне показался данный подход достаточно зрелым, поэтому я взял его за основу и попробовал упростить для того, чтобы сделать единую методологию.

Прежде стоит отметить, что раньше аудиторам помогали специальные напарники, которые аудировали, как работают IT-системы, хранящие первичную информацию для отчетности.

Эти люди при проверке IT-систем изучали, как работает контроль в отношении данных. Должны были быть ответы на такие простые вопросы: «Откуда данные?», «Кто может их изменить?», «Как проверяется корректность значений?», «Какие программные средства использует организация для исправления проблем?» и так далее.

Они используют опросы, изучают логи[133] подключений к системам и на выходе, по результатам своей работы, они могут сказать, в каких из измерений, скорее всего, будет проблема.

Эти самые «assertions» можно смело назвать «измерениями», то есть некоторым разделением того, как я воспринимаю объект в реальном мире.

Главное, что они должны говорить пользователю – любое число или любые данные – само по себе объект многомерный.

Вот я держу книгу. В стандартной проекции у нее три оси –

ширина от края разворота до середины, длина от одного края страницы до другого края страницы и толщина, то есть количество страниц. Книгу мы воспринимаем как физический объект в трех измерениях.

Так вот, информация сама по себе имеет много измерений, больше трех. И не факт, что их именно тринадцать. Чтобы управлять качеством этой информации, нужно управлять представлением этой информации в этих измерениях. Это сложный контекст, отчасти поэтому в качество данных мало инвестируют и мало этим занимаются, хотя, на мой персональный взгляд, ценность этого очевидна.

Чтобы стало проще, можно упростить количество тех самых измерений, в которых мы управляем качеством данных. Для простоты оставим только «полноту» и «точность» – то есть все, что произошло вокруг, отражено в информации и отражено корректно. Только два измерения.

Теперь вернемся к пресловутому и коварному отчету «аппетит к риску» – здесь мы должны посчитать размер потенциального искажения для двух измерений.

Как пострадает организация, если поймет, что не отражены только 95 % тех событий, которые произошли, или что сами 15 % событий отражены неточно? Возьмем то же поле «ИНН». Допустим, что поле заполнено только в 95 % случаев, а в заполненных оно некорректно в 15 % случаев. Пусть мы говорим о количестве записей 10 тысяч единиц известных нам, тогда потенциальный размер штрафа будет равен:

15 %*95 %*10 000 + (10000/95 % – 10000) = 1425 + 526 = 1951 записи могут быть некорректны.

Опустим как получили оценку 95 % или 15 %, для простоты считаем это экспертной позицией участников процесса работы с данными.

1951 умножаем на размер штрафа в пятьсот рублей, получаем 975 500 рублей – это потенциальный убыток от проблем с качеством данных одного поля «ИНН» для организации.

Как понять, какие измерения качества выбрать?

Мне нравится одно очень интересно исследование, которое провели исследователи из MIT. Оно называется «**Beyond Accuracy**»[134][135]. Для него исследователи выделили несколько групп пользователей.

В первой группе пользователей, которых они опросили, были студенты MBA.

Во второй группе пользователей, среди которых был опрос, находились уже выпускники MBA, которые проработали в компаниях достаточное количество лет.

Опросы так же отличались друг от друга. Опрос для первой группы включал в себя список возможных измерений контроля качества данных, из которого студенты должны были выбрать предпочтительный.

Напомню, что в исследованиях всегда используют один из трех различных подходов получения научного познания:

• Эмпирический – познание получаем через ощущения.

• Теоретический – осмысление опыта с точки зрения логики.

• Интуитивный – когда мы полагаемся на свой «внутренний» голос при исследовании того или иного события.

В первой группе исследователи применили теоретический подход получения нового знания – а именно списка параметров, «измерений», по которым можно контролировать качество данных.

Во второй группе исследователи применили уже интуитивный подход, чтобы понять, какие из этих параметров на самом деле наиболее важны в принятии решений и их влиянии на бизнес. В этом случае продолжительный опыт бывших выпускников MBA в компаниях являлся тем самым «внутренним фильтром», который помог определить наиболее ценные измерения из большого списка.

Исследователи сформировали список из 32 параметров контроля качества данных (32 параметра – это достаточно внушительно), и попросили сформулировать, как бы выпускники контролировали качество данных.

По итогам опроса получилось 179 уникальных параметров, которые сформулировали участники процесса, то есть в пять с половиной раз больше, чем исследователи изначально заложили в свою модель.

Модель исследователей строилась на четырех основных

группах, которые объединяли эти самые параметры:

• Доступность – данные должны быть **доступны** для пользователя.

• Интерпретируемость – данные должны быть способны к **интерпретации**. К слову, не пытайтесь использовать мандаринский диалект, если вдруг пишите комментарии в проводках и так далее.

• Релевантность – данные должны быть **релевантны** для конечного пользователя, если они участвуют в процессе принятия решения.

• Точность – данные должны быть **точны** для пользователя, то есть быть точными и из достоверных источников.

Во второй группе исследователи отбросили часть новых параметров и показали только 118 параметров контролирования качества данных. Опрос строился на ответах 1500 выпускников MBA, которые уже имели внушительный опыт работы.

Опустим тот факт, что опрос строился через почту, и тогда не было еще нормального работающего Интернета, обратимся лучше к его результатам.

99 из указанных параметров из основного списка оказались абсолютно не важны, когда люди с большим опытом и багажом знаний попытались интуитивно ответить на тот же самый вопрос о том, как контролировать качество данных.

Два параметра пользователи выделили как самые важные – «точность» (accuracy) и «правильность» (correct). Все самые важные параметры исследователи сгруппировали вместе в кластеры, которых получилось ровно четыре.

Качество данных
Внутреннее качество данных (Intrinsic DQ)
Качество данных контекста (Contextual DQ)
Качество представления информации (Representational DQ)
Доступность информации (Acessibility DQ)

*Рисунок SEQ Рисунок * ARABIC 1 Структура концептуального фреймворка DQ, на основании исследования MIT Beyond Accuracy. 1993*

Внутреннее качество данных – включает не только точность, но и два новых измерения – репутацию и

правдоподобие. Одна лишь «точность», как оказалось, не дает пользователям уверенности в корректности данных. Им нужно доверять источникам данных.

Качество данных контекста – как оказалось, качество данных по контексту профессиональная литература по работе с данными не распознает, то есть, таких знаний просто не было. Люди не имели представления, как управлять качеством того контекста, который они получают. Единственные доступные материалы были о качестве визуального контекста – графике. Мы это подробно разобрали в главе про «Data Storytelling». Пример реализации контекстных проверок был, как ни странно, в армии Соединенных Штатов Америки во время операции «**Буря в Пустыне**»[136], где такие проверки были установлены на воздушных судах. Они анализировали для каждой задачи, выполняемой воздушным судном, широкий список параметров, используемый в планировании авиаударов.

Представление качества данных – в первую очередь эта группа касается проблем с форматом данных и с тем, чтобы данные можно было понять и интерпретировать. К примеру, данные по отчетности ВТБ отражаются в российских рублях, в свою очередь, в данных группы Альфа-Банка в публикуемой отчетности вместо рублей уже используются доллары как основная валюта.

Доступность данных – один из самых неоднозначных параметров, потому что управление информационной безопасностью в большинстве проектов и организацией, в которых мне довелось побыть, выведено за периметр как IT-департамента, так и Финансового департамента. Управление информационной безопасностью – это отдельно выделенный лидер внутри организации, поэтому решения в области ИБ в большинстве случаев не участвуют в управлении качества данных.

В итоге, исследователи MIT вывели 15 ключевых измерений того, как можно управлять качеством данных, и сформировали из них следующий фреймворк.

Target category	Dimension
Intrinsic DQ (Accuracy of data)	Believability Accuracy Objectivity Reputation
Contextual DQ (Relevance of data)	Value-added Relevancy Timeliness Completeness Appropriate amount of data
Representational DQ (Representation of data)	Interpretability Ease of understanding Representational consistency Concise representation
Accessibility DQ (Accessibility of data)	Accessibility Access security

Это было более двадцати лет назад, но на мой скромный экспертный взгляд такой подход по-прежнему актуален, хотя мало где еще используется. Его мы можем смело использовать при подготовке отчета «аппетит к риску», чтобы выбить из менеджмента ресурсы на все свои «хотелки» в области данных.

Инструменты управления качеством данных

Вот представим, что вам дали задачу наладить контроль качества данных в организации. Ваши действия? Кроме того, что выпить валерьянки – это и так понятно.

Я бы постарался разделить весь этот необъятный пирог из данных внутри организации на какие-то разумные блоки или куски.

Интуитивно понятным мне видится выделить хотя бы три блока информации, которыми можно попробовать управлять каждым по-своему. Ими будут – клиенты, справочники и продукты.

Такие блоки для простоты предлагаю называть «доменами», только не будем путать их с теми доменами, которые есть в Интернете. В текущем контексте «домен» – это группа однородной информации, которой нужно управлять. Большую часть информационного пирога можно разделить на три крупных блока, чтобы с этим начать работать по закону Парето[137].

Почему так?

На моей практике оказалось, что решения или инструменты (программные средства), которые обещают стать универсальным средством, по факту проигрывают в этом соревновании. Либо они становятся малоэффективными, то есть не позволяют выявить проблемы в данных, либо они становятся неимоверно дорогими, до такой степени, что стоимость их использования совершенно не сопоставима с получаемой ценностью.

Поэтому я для себя решил, что одного универсального средства ото всех бед не существует. Кстати, такие универсальные инструменты называются MDM-платформами[138]. Какое-то время они считались единственным средством против всех болезней и рекомендовались к внедрению в любой организации для любой проблемы. В реальности каждое такое внедрение превращалось в некий эпик фейл, то есть в мероприятие, обреченное на провал. Поиск Святого Грааля в решениях с данными натолкнул на мысль о «вырожденности» решений, которые могут управлять различными доменами. «Вырожденность» подразумевает, что свойства и функции каждого из инструментов для различных доменов сильно отличаются друг от друга. Инструмент по управлению качеством данных для домена «Клиенты» не подходит для управления качеством в домене «Справочники» – и наоборот.

Теперь шаг в сторону и маленький ликбез по тому, как можно управлять качеством данных этого большого информационного пирога.

С одной стороны, можно не трогать источники данных и работать с конечным информационным продуктом, что обычно получается на выходе. К примеру, как это делают аудиторы, когда работают и проверяют финансовую отчетность и данные, на основании которых она строится.

С другой стороны, можно исправлять данные там, где они появляются, то есть, в самих системах. Или исправлять до того момента, пока они не появятся. Например, ограничить ввод данных и задать определенные рамки для информации[139].

Разница между этими принципиальными подходами заключается только в стоимости усилий. Оказалось, что стоимость усилий контроля в конце цикла какого-то длинного процесса **кратно** выше, чем стоимость контроля на начальных этапах. Происходит это из-за того, что анализ проблем осуществляется в конце, и требуется много дополнительного времени для раскопок причин, которые обычно могут рассказать, что пошло не так.

Если, скажем, стоимость проверки в конце при формировании финансовой отчетности стоит тысячу рублей, то стоимость контроля на первых этапах будет стоить не больше десяти рублей. Поэтому при плохой работе внутренних контролей внешний аудит обычно стоит много денег, потому что аудиторам приходится раскапывать много информации, – почему все плохо в цифрах.

Работать с качеством данных и контролировать их на первых этапах можно несколькими путями. Можно задать тот самый коридор допустимых значений, которые, к примеру, не позволяют ввести несуществующий адрес.

Во втором случае все работает наоборот. Решения допускают ввод любого значения, чтобы потом некоторое внутреннее решение по тому или иному домену само разобрало данные с использованием различной сложной логики и предложило правильный вариант. Если я ввел несуществующий адрес, система сможет найти самый ближайший аналог или экземпляр похожего адреса, который существует в реальности, а дата-стюард уже согласует финальное значение.

Да, именно так, решения сами по себе без дата-стюардов не работают, то есть нельзя автоматизировать всеми возможными средствами вокруг все возможные ошибки.

Самыми эффективными считаются гибридные подходы. Они объединяют первые подходы с заранее невозможным вводом

несуществующих и невозможных данных и с определенным допущением свободны со стороны пользователя ввести все, что он считает правильным. Пример с почтовыми индексами это наглядно отражает. В базах данных они по-прежнему некорректные, поэтому пользователя просят дополнительно в свободной форме ввести его.

Первый подход называется «децентрализованным» контролем качества данных[140], а второй, когда задаются значения, называется «централизованным»[141].

Домен «Клиенты» – вся информация, которая касается наших клиентов: их ФИО, дата рождения, контактные данные, сегменты, в которые определил их маркетинг, выводы, которые сделал комплаенс, и так далее. Все это будет внутри домена «Клиенты».

Для управления качеством данных в этом домене используется специальное решение CDI[142], задача которого посредством специальной сложной логики уметь сопоставлять различный образцы карточек клиентов, выделять похожих и указывать на это дата-стюарду.

Как было в случае «склейки» меня и моего брата в банковских сервисах, такое решение должно было выявить ошибку и показать дата-стюарду, что две карточки клиентов с одинаковой фамилией, одинаковым отчеством и одинаковой датой рождения склеились, но на самом деле это разные клиенты, так как у них разные имена и паспортные данные.

Правила в CDI задает и проверяет никто иной как инженер данных. Надеюсь, теперь стало понятно почему эти ребята тоже крайне важны.

Как работает CDI?

Он превращает информацию каждого экземпляра карточки клиента в сложный код посредством определенной логики и сравнивает их между собой. Например, внутри банка может такое быть, что Благирев Алексей Павлович был заведен девять раз в различных системах, и данные, естественно, неоднородно заполнены между всеми этими системами.

Где-то нет даты рождения, где-то нет полных паспортных данных, где-то нет адреса и много чего другого.

CDI объединяет все эти образцы вместе и создает свой собственный уникальный образец, который уже включает в себя все заполненные параметры из различных источников. Этот конечный образец называется «золотая запись», его можно уже передавать в системы и использовать.

CDI позволяет организовать полноценный промежуточный слой работы с клиентскими данными, а на него уже можно

«надеть» или подключить все основные клиентские сервисы через CRM.

Но встает вопрос – как быть уверенным, что система взяла нужный образец и вытащила нужную дату рождения для этого образца? Здесь как раз снова появляются инженеры данных, которые определяют допустимые критерии (веса) по тем или иным источникам данных, полям и так далее. В определении весов участвует как интуиция, так и статистика. Сколько существует однофамильцев, которые родились со мной в один день? Инженеры знают ответ. Ну или должны знать.

Домен «Справочники» – тут нужно разобраться, какие из доступных классификаторов внутри организации являются ключевыми, то есть такими, которыми пользуются все департаменты. Эти классификаторы можно назвать глобальными. Классификатор – это список допустимых значений – названия офисов, список продуктов, список сегментов и так далее.

В качестве технического решения используется специальное средство RDM[143] или по-русски «НСИ»[144], которое не просто хранит правильный список значений и его распространяет, но и имеет встроенный механизм управления изменениями этих значений. Этот механизм допускает ввод новых значений только от владельцев данных.

Да именно так, появляются владельцы данных, которые отвечают за корректность того или иного справочника.

Владельцы могут назначаться на конкретный параметр в справочнике, а сам справочник может быть представлен не просто списком, а довольно сложной иерархией (отделения складываются в филиалы, филиалы складываются в организацию).

Рассмотрим пример со справочником банковских офисов. У него должен быть определен владелец, который отвечает за качество значений всех офисов. Должна быть процедура заведения нового значения в справочник.

Например, нам нужно поставить в банковский офис пандус, чтобы люди с ограниченными возможностями или дама с ребенком в коляске могли зайти в офис. Но так, чтобы дама сразу знала какой из офисов оборудован пандусом, а какой нет. Для этого руководителя офиса можно сделать владельцем данных одного параметра в справочнике банковских офисов – «Есть пандус».

Руководитель банковского офиса, который непосредственно находится на месте будет отмечать галочку «Да»/Нет», если в офисе есть пандус, а система уже выведет эту информацию на сайт или в мобильное приложение, чтобы конечный пользователь

смог выбрать ближайший к нему офис и без проблем заехать в него.

Домен «Продукт» – самый сложный на мой взгляд домен, его цель – управлять жизненным циклом продукта внутри организации. От момента его создания, до момента его снятия с производства или с продаж. В розничном бизнесе и банках такие IT-платформы, которые управляют качеством данных по продукту называются PIM[145]. В первую очередь, это управление каталогом продуктов и характеристиками каждого из продуктов, сбор статистики и определение базовой себестоимости услуг и сервисов внутри каждого конкретного продукта. На производствах такие платформы более комплексные, так как там необходимо уже интегрировать много различных источников (3D схемы из CAD решений и другие), они называются PLM[146]. Они содержат информацию об изделии: 3D схему, технологическую карту о том, как изделие изготовлено, технологический паспорт и инструкцию по ремонту, то есть как изделие необходимо обслуживать.

На практике очень мало результативных внедрений таких технологий, потому что они затрагивают множество процессов в организации и являются критерием зрелости компании в работе с данными.

Начинать с домена «Продукт» я не рекомендую, потому что домены «Клиент» и «Справочники» являются гигиеническим минимумом в организации. Если компания решилась всерьез заняться качеством данных, переход к «Продукту» будет наиболее эффективен, когда организация освоит не только сами инструменты RDM или CDI, но и запустит соответствующие службы из дата-стюардов и дата-инженеров.

Глава 7
Не Big Data единой: платформы и экосистемы

PaaS и платформы

Платформа – это бизнес-модель, которая позволяет нескольким участникам (производителям и потребителям) подключаться к ней, взаимодействовать друг с другом, создавать и обменивать стоимость[147].

Примеры успешных компаний, реализовавших платформенную модель – Alibaba, Amazon и Facebook.

Что такое DMP-платформа?

При работе с цифровым маркетингом возникает вопрос: как сделать так, чтобы предложение максимально соответствовало потребностям клиента? Продать случайному человеку случайный товар из ассортимента гипермаркета сложно. Это как стрелять в небо. Но если мы знаем, что человеку необходимо или понадобится в ближайшем будущем, шансы на успешную продажу значительно вырастают.

Например, продать автомобилисту зимой после снегопада новую щетку для снега – труда не составит. Но как узнать, у кого есть машина, а кто из владельцев автомобилей давно не покупал новую щетку?

Где взять такие данные? Каким образом их обработать?

На помощь приходят технологии сбора и анализа Больших данных.

А происходит это так.

Любое действие в Интернете оставляет в сети след. У каждого пользователя (на самом деле у браузера пользователя) есть свой уникальный идентификатор – cookie. Он позволяет отслеживать действия пользователя на сайте (или нескольких) в течение ограниченного количества времени. Затем идентификатор меняется.

Если пользователь находится на сайте или заходит на него через разные браузеры, то у него будет много разных идентификаторов. Cookie также могут устанавливать внешние серверы, не имеющие срока окончания. Это позволяет отследить сессии пользователя при повторном появлении в Интернете. При этом, cookie не всегда точно определяют тип пользователя, потому что за компьютером могут находиться попеременно разные люди.

Многие сайты устанавливают специальный код, который называется «пиксель». Назван он так потому, что загружается в

виде изображения с минимальными размерами один пиксель на один пиксель. Его задача – собирать данные о посетителях сайта, включая их cookie.

Что происходит дальше?

Пиксель передает данные об аудитории в единую платформу. Так, данные с пикселя Facebook, который установлен на сайте, передаются в единую платформу обработки данных Facebook. Платформа определяет, к каким частям (сегментам) аудитории относится клиент, и таким образом запоминает его характеристики.

Сегментировать аудиторию можно по интересам, возрасту и так далее. Для этого используются самые разные методы. Самый простой метод – «если – то»: если пользователь пришел с сайта о спиннингах, то ему может нравиться рыбалка. Метод сложнее – машинное обучение.

Так вот, такие платформы, которые собирают данные с сайта и позволяют обогатить их дополнительной информацией, называются DMP (Data Management Platform или «платформа для управления данными»).

Данные, собранные с пользователей, можно использовать. Например, сделать маркетинговую кампанию точечной, коммуницировать только с подходящей под нее группой пользователей:

– предложить существующим клиентам какой-то продукт в дополнение к действующим продуктам.

– предложить новым пользователям уникальные условия, чтобы они стали новыми клиентами.

Таким образом, данные, получаемые из DMP через пиксель, направлены на повышение эффективности конверсии, то есть на превращение новых пользователей в клиентов, клиентов в лояльных клиентов и так до бесконечности.

DMP1.0

История DMP начинается с ростом популярности цифрового data-driven маркетинга. В нем предложения строятся на основе анализа данных о продажах, клиентах и, часто, не связанных с ними напрямую данных.

Цифровой маркетинг был представлен несколькими уровнями, на каждом из которых работают соответствующие платформы:

– DMP (Data Management Platform) – цифровая платформа, которая собирает, агрегирует различные данные о пользователях,

чтобы получить информацию о сегментах аудитории.

– DSP (Demand Side Platform) – цифровая платформа для покупки баннерной, видео, мобильной или поисковой рекламы. Данная платформа исполняет код и предлагает пользователям те или иные товары на основании подготовленного ранее профиля со стороны DMP, то есть она управляет кампаниями. DSP представляет покупателя рекламного контента.

– SSP (Supply Side Platform) – инструмент измерения монетизации посещений цифровых ресурсов (веб-сайта или мобильного приложения), которые он оптимизирует для издателей (publisher) предложения по рекламным позициям для рекламных бирж (AdExchange), подкрепленные результатами по анализу эффективности конверсии. SSP представляет продавца рекламного контента.

– AdExchange – рекламные биржи, которые подобно фондовым биржам обрабатывают, размещают рекламные объявления и устанавливают взаимоотношения покупателей и продавцов (рекламодателей – advertisers, тех, кто публикует рекламу).

Например, если компания Coca-Cola захочет запустить кампанию в Интернете, то она обратится к соответствующим медиа-агентствам, те, в свою очередь, разместят заявку на проведение кампании на соответствующей рекламной бирже, а ее уже увидят сервера рекламных агрегаторов и покупателей.

Одна из таких первых бирж была открыта в США в 1996 и называлась DoubleClick. Вначале она работала как application service provider, то есть предоставляла сервисы и приложения для пользователей в виде рекламы. В 1999–2001 годах DoubleClick, обслуживающая и предоставляющая рекламу на тот момент для 11,5 тысяч веб-сайтов, провела серию поглощений компаний, став крупнейшей онлайн-биржей по рекламе. Были куплены несколько быстрорастущих компаний NetGravity и Abacus Direct. Позднее, в 2007 году, Google купил DoubleClick за 3,1 миллиарда долларов, организовав на ее основе полноценную маркетинговую платформу.

В реальном времени покупатели и продавцы взаимодействуют друг с другом для покупки и размещения рекламы в digital-каналах – на сайтах, в мобильных приложениях и поисковых запросах. Основу этого взаимодействия выстраивают как раз данные, собирать которые помогают DMP-платформы. Но DMP-платформы обычно только представляют сторону покупателя (рекламодателя).

Они собирают три типа данных:

– Собственные данные (First party data) – данные

пользователя, в том числе персональные и контактные, а также информация о его действиях в digital-канале (мобильное приложение, сайт, поиск и так далее), все это компания собирает сама.

– Данные о маркетинговых активностях (Second party data) – результаты проведения кампаний, информация по откликам, конверсии, а также поведенческих факторах, которая компания может купить у других источников, где был пользователь. Этакий аналог собственных данных, которые купили у другой компании.

– Данные третьих лиц (Third party data) – сегменты и аналитика, которую предоставляют специальные провайдеры данных, благодаря синхронизации через cookie с различными поставщиками данных.

Отличие DMP от CRM-систем

Часто маркетинговые кампании проводятся с участием CRM (Customer Relationship Management – системы управления взаимоотношениями с клиентами). Поэтому DMP и CRM иногда путают. Однако, DMP и CRM не равны. Маркетинговая компания InBrief выделила ключевые отличия платформ друг от друга:

DMP используется в основном для привлечения новых клиентов и расширения уже существующей клиентской базы за счет look-alike моделирования (то есть поиска групп клиентов, похожих на существующие).

На DMP могут собираться все типы данных.

Персональные данные внутри системы максимально отделяются от остальных. Это делается из-за законодательных ограничений.

DMP разрабатываются по большей части для взаимодействия с рекламными сетями.

Принцип работы заключается в поддержании большой cookie-базы с анализом уже «встроенных» аудиторий, чтобы найти максимальное соответствие для рекламной или маркетинговой кампании.

CRM используется преимущественно для удержания потребителей и развития уже существующей базы клиентов (например, для увеличения активности во время промо-акций, числа покупок в чеке и так далее).

Собираются преимущественно first- и second party данные.

Потребителям присваивается уникальный ID, создается и дополняется персональный профиль на основе множества каналов (известных и анонимных), доступ к которому можно использовать

для различных целей.

Разрабатывается преимущественно для сбора данных за счет интеграции маркетинговых каналов.

Принцип работы заключается в механизме персонализации с анализом поведения, вовлечения в контент и так далее, в целях дальнейшей предиктивной аналитики (например, где вероятнее «отток» клиентов) и еще большей персонализации.

DMP 2.0 – Траектория развития

Согласно данным компании IDC, ежегодный объем прироста мировых данных составляет 16.3 зеттабайта (триллион гигабайтов). К 2025 году этот прирост достигнет отметки в 163 зеттабайта. Тогда будут доступны новые данные для анализа, управления и расчета наиболее подходящего предложения.

Среди новых групп данных есть, например, психографика, нейрофизиология, данные с бытовых устройств (умные холодильники и прочие приборы). Психографика позволит делать умную персонализацию – например, в письме от интернет-магазина будет не только ваше имя, но и привычные вам стиль и лексика.

Умные устройства смогут поставлять много (относительно) честной информации о поведении пользователей. Если ваш умный горшок для цветов уже полгода подает сигналы о том, что цветы в нем засохли, и сайты про садоводство вы больше не посещаете, то рекламу с новыми товарами на эту тему вам, вероятно, могут больше не присылать.

Все это сильно повлияет на существующий ландшафт решений по предоставлению цифровых товаров.

Для определения существующих векторов развития следует разобраться с развитием платформ, управляющих клиентскими данными:

– Эволюция CRM → ACRM → CDP (Customer Data Platform)

– Эволюция DMP → DXP (Data Exchange Platform)

Эволюция CRM

Большинство бизнес-приложений имеют внутри встроенную, либо подключенную извне (как например Salesforce.com) платформу CRM для управления данными о своем клиенте.

Следующим этапом идут платформы автоматизации предложений под те или иные сегменты аудитории. Несмотря на

то, что инструментов на рынке много, примеров успешной интеграции платформ Marketing Automation (они же ACRM) мало. Но все же компаниям требовалось научиться объединять данные пользователя о его действиях в сети, его учетной записи и, например, информации о его взаимодействии с call-центром компании.

В этом смысле CDP является большой базой, которая объединяет как данные, получаемые из фронт-офисных систем (CRM), так и данные из третьих сторонних источников (DMP), но, в отличии от DMP, CDP работает с профилем, а не только с cookie.

Сегодня в основе успешного CRM лежит CDI (Customer Data Integration). Это категория платформ работы с персональными данными, которая позволяет унифицировать профиль клиента, работая на уровне контроля качества данных и объединяя похожие образы следов или профилей клиента.

Но, в отличии от CDP, CDI работает со внутренними системы компании, получая единый профиль клиента.

Если работать над объединением offline-online данных за пределами систем компании, то в таком случае CDI может эволюционировать в CDP для решения задач по получению единого профиля.

Отличительная особенность CDP – это поддержка автоматизации маркетинговых кампаний.

Эволюция DMP

Первые DMP были созданы до того, как Google/Facebook открыли свои API для работы с аудиторией (устанавливать пиксель, работать с сегментами и другие вещи, о которых было рассказано выше).

С появлением этих больших игроков система устоялась и стала основной инфраструктурой для развития цифрового маркетинга. Но в ней есть недочет: возможность продавать рекламные объявления есть, а вот возможности продавать и покупать данные на уровне сервиса – нет.

Именно поэтому в качестве следующего шага развития DMP-платформы будут превращаться в DXP (Data Exchange Platform), в так называемые биржи по обмену данными, где можно разместить свои данные или купить необходимые, заключив прямой контракт с их владельцем.

Data Exchange Platform еще не имеет устоявшегося определения.

Например, по версии исследовательской компании Gartner,

такие платформы называются Digital Marketing Hub.

Известный эксперт в области цифрового маркетинга Джонатан Бистон разобрал магический квадрант Gartner для DMP-платформ. На нем были указаны сразу и DSP-платформы, и DMP, и SSP, и это сильно сбивало всех с толку. Используя свой уникальный опыт по работе как в AdTech, так и в компании Adobe, он переделал этот магический квадрант в более подходящую форму, способную раскрыть суть Digital Marketing Hub, объединившего сразу несколько направлений.

Основная мысль, которую хотел донести Бистон складывается так:

«Взгляд на то, что такое Digital Marketing Hub, зависит от перспективы, с которой смотрит основной потребитель, входящий в одну из четырех групп. Каждый из четырех потребителей видит конечную цель исключительно в своей плоскости, но при этом инфраструктура может быть общей и совмещать миллиарды новых данных о пользователях, устройствах и используемом контенте».

Экосистемы

Экосистема представляет собой набор услуг, который объединяет пользовательский опыт.

Потребительские экосистемы, как правило, сосредоточены на таких потребностях как путешествия, здравоохранение и жилье. Системы B2B обычно вращаются вокруг определенного лица, принимающего решения, например, маркетинга и продаж, операций, закупок или профессионалов в области финансов.

В чем преимущества экосистем?

– Они действуют как шлюзы и снижают затраты клиентов на переключение на связанные услуги. Например, мессенджеры WeChat и Line позволяют пользователям совершать покупки, регистрироваться на мероприятиях, читать новости и общаться с врачами через один интерфейс. Пользователям не нужно переключаться между порталами, управлять множеством паролей и вообще помнить про несколько сервисов.

– Они используют сетевые эффекты. Google Nest стал основой для экосистемы смарт-домашних продуктов и, например, присылает своим клиентам ежемесячную карточку отчета по использованию энергии и сравнивает этот расход с показателями соседей (и это дает контекст). Одновременно с этим компания передает агрегированные данные поставщиков коммунальных услуг. Эта информация может помочь им оптимизировать свои

процессы.

– Они объединяют данные по ряду услуг. Одна медицинская компания извлекает данные высокой точности из системы здравоохранения и применяет ее к жизни пациентов для улучшения здоровья человека. Другой пример: сервис Dash берет данные по отзывам и сервисным компаниям у автопроизводителей, а потом делает персонифицированные рассылки своим клиентам.

Консалтинговая компания McKinsey прогнозирует появление двенадцати глобальных экосистем, относящихся к различным сферам бизнеса, к 2025 году.

Глава 8
А что дальше? Проблемы и тренды

В 2015 году исследовательская компания Gartner убрала Big Data со своей «кривой хайпа». Но до сих пор вокруг этого термина существует какая-то лихорадочная активность. По-прежнему идет речь о Big Data-трансформации, но далеко не всегда понятно, что это такое, и какую конкретно пользу оно может принести бизнесу. Сам по себе переход на новые технологии вряд ли может привести к увеличению прибыли или сокращению накладных расходов.

Проблемы с Big Data сегодня

Хотя технологии Big Data сейчас уже применяются промышленно, бóльшая часть проектов в этой области не имеет успеха. Почему?

Мы думаем, что понимаем Big Data

Проекты, связанные с Big Data-аналитикой, часто воспринимаются всеми (менеджментом и самими разработчиками) как традиционные IT-проекты с фиксированным скоупом (объемом работы).

В реальности же это, скорее, RnD-проект (Research and development или исследование и разработка). И ключевую роль здесь играет именно исследовательская часть. На самом деле, не определены ни конечный результат, ни время, за которое будет получено хоть что-то.

Big Data аналитика – это постоянное исследование, в ходе которого скорее появятся внезапные полезные инсайты, чем стабильные и быстрые бизнес-результаты (конечно, если речь идет о новом проекте). Однако то, как раскрывается ценность этих инсайтов, зависит больше от знания предметной области, чем от количества данных, математической или технической сложности решения. И здесь как никогда справедлива фраза «отрицательный результат – тоже результат», только надо уметь это увидеть.

Еще одна проблема – недостаток специалистов. Покупка инструментов и применение agile-методологии в полной мере ее не решает. Уровень опыта и экспертиза также играют роль в успешном завершении исследовательских проектов Big Data.

Как рассчитать финансовый эффект?

Большая гибкость в отношении сроков и результатов проекта ведет к необходимости выделения большего количества ресурсов. Оно начинает слабо и предсказуемо расти, когда компания сталкивается с реальными долгосрочными задачами и необходимостью соблюдать SLA, а также требования регуляторов.

Сроки гибкие, результат непредсказуем – значит, на проект может уйти больше ресурсов (времени, людей, денег), чем предполагалось.

Проекты, связанные с Большими данными, не всегда решают уникальные задачи. Эти проекты считаются научными без каких-либо бизнес-целей или показателей. Чтобы извлечь максимальную выгоду из этого, нужно направить усилия на конкретную потребность или проблему бизнеса. Чтобы оправдать инвестиции для проектов Big Data, требуется постоянно демонстрировать результаты. Бизнес требует быстрого и гибкого доступа к данным с прозрачными SLA. В результате оказывается, что бизнес ожидает большого количества дешевых инсайтов, а Big Data- и Data Science-специалисты требуют ресурсов на исследовательскую составляющую проектов и большую толерантность в ошибках и неудачах, являющихся неотъемлемой частью их работы. При правильном использовании, Big Data дает широкий спектр возможностей для бизнеса сегодня и в будущем. Проблема заключается в нехватке квалифицированных специалистов и неравномерной выдаче результатов. Это только вопрос времени, когда Big Data станет важной частью принятия бизнес-решений. Если эти ошибки будут учтены, станет намного проще реализовать любую стратегию, связанную с Большими данными. Еще один способ увеличить шансы на успех – использовать правильные инструменты для правильного проекта.

Вообще, все бизнес-цели можно разделить на два больших направления:

– Создание и запуск нового сервиса с использованием данных

– Оптимизация текущего процесса или сервиса с использованием данных

На практике необходимо забыть о сложности самих технологий и ограничениях в компетенциях, и использовать все возможности как необходимые компоненты при трансформации.

В первом случае расчет доходности использования технологий ничем не отличается от расчета окупаемости инвестиционных затрат при запуске нового продукта. Как ни банально, но мы переступили черту, где хотели кого-то удивить,

и попали в мир, где уже «так принято» использовать данные.

Во втором случае финансовый эффект можно оценить по той части процесса, которую мы собираемся менять. Сравнивается себестоимость текущего звена процесса, размер сопутствующих операционных рисков и инвестиционных затрат на разработку и замену этого компонента сервисом с использованием данных. Строится описание текущего процесса, который планируется затронуть с использованием одной из существующих общепринятых нотаций (EPC[148], BPMN и других), где нужно заменить один из типизированных этапов в процессе:

• **Ручной ввод** со стороны человека заменяется обработкой и анализом ранее введенных логов. Большинство значений внутри процесса стандартизируется и классифицируется, и человек вводит данные только в исключительных случаях. При этом алгоритм может запоминать введенное значение, чтобы его не нужно было вводить снова. Сравнивается себестоимость текущего звена процесса, размер сопутствующих операционных рисков и инвестиционных затрат на разработку и замену этого компонента сервисом с использованием данных. Такие сервисы в среднем окупаются за срок не более полутора года. Бывают случаи, когда сервис не может полностью заменить оператора процесса: например, оператора для чата мобильного банка, взаимодействующего с пользователем. В этом случае сервис может вместо набора текста использовать всплывающие подсказки, если ему не удастся полностью распознать запрос.

• **Сверки и реконсилиации –** это целый этап в бизнес-процессе, на котором пользователи тратят время на проверку и аудит полученной ранее информации. Его можно заменить на автоматические проверки. Например, в процессах выпуска финансовой отчетности есть очень много точек, когда информация проверяется и сверяется с источниками данных перед тем, как попасть на стол к финансовому директору.

Big Data может быть вообще не нужна

Big Data – это модная и современная технология, и часто возникает соблазн везде ее использовать.

Причины могут быть разные.

– Когда в руках молоток, все вокруг – гвозди.

– Незнание предметной области.

– Необходимость произвести впечатление на бизнес и публику. Было время, когда стартапы, не использующие «ML/AI» просто не воспринимались всерьез.

– Просто интересно попробовать новое. И в этом нет ничего плохого, если отдавать себе отчет о сроках, ресурсах и возможных последствиях.

На вопросы бизнеса зачастую можно ответить с помощью простого SQL. А бизнес-логику сделать на нескольких сценариях «если – то».

И все же, несмотря на эти проблемы, технологии продолжают развиваться и двигаться вперед.

К чему мы движемся? Тренды

Облачные решения

В 1980–х появилась концепция Plug and Play (англ. включил и играй/работай). Она позволяла собрать свой домашний компьютер из отдельных деталей, у которых были стандартные интерфейсы. Облачные технологии позволяют сделать то же самое, но уже по отношению к бизнес-процессу или бизнесу в целом. Концепции Process-as-a-Service, Data-as-a-Service, Analytics-as-a-Service уже сегодня позволяют собрать как из конструктора работающую IT-систему для бизнеса.

Облака позволяют сократить затраты на инфраструктуру и ее обслуживание. Еще одно преимущество – быстрое масштабирование. При возросшей нагрузке мы можем быстро увеличить количество доступных системе ресурсов.

Ожидается, что к 2020-му году (по крайней мере) треть всех данных будет проходить через облако.

Лидеры рынка, которые способны эффективно анализировать несколько источников данных, могут использовать различные возможности для повышения эффективности работы. Крупный бизнес уже начал активно менять свои процессы и переносить данные и работу с ними в облако.

– Вся инфраструктура Pinterest находится в облаке.

– Компания Xerox использовала стратегию облачных вычислений для эффективного анализа данных и снижения скорости изнашивания в своем call-центре на двадцать процентов.

– Компания Caterpillar разрабатывает специальные облачные решения для анализа и отслеживания того, как работает ее техника в совокупности с предоставляемыми финансовыми сервисами, что позволяет существенно сократить расходы на аудит и мониторинг объектов, которые могут быть заложены в рамках сделках финансового лизинга.

– Компания Боинг в 2015 году перешла на облачную платформу. Это ускорило более чем в 100 раз работу ее служб доставки и в шесть раз увеличило утилизацию **активов**[149].

В России бизнес не всегда спешит переходить на облачные технологии. Это связано с тем, что большинство крупных облачных провайдеров – это зарубежные компании. Поэтому возникают законодательные ограничения и риски, связанные, например, с курсом валют.

Кроме того, многие традиционно не доверяют третьим лицам

данные, представляющие собой коммерческую тайну.

Тем не менее, облачные решения появляются и на нашем рынке.

Машинное обучение применяется все чаще

По мере того, как развивалась Big Data-аналитика, некоторые компании стали инвестировать в машинное обучение (ML). Машинное обучение остается одной из самых востребованных и внедряемых технологий. И она еще не исчерпала свой потенциал. По прогнозу аналитической фирмы **Ovum**[150], машинное обучение – один из главных трендов в Big Data-технологиях. Его применение будет все расширяться. От задач по бизнес аналитике оно перейдет на большинство задач по подготовке и преданализу данных. Не исключено, что ряд задач по интеграции источников данных также будет решаться с привлечением машинного обучения через анализ и интеграцию словарей (описание объектов данных в тех или иных источниках).

Аналитика всего

Предсказательная аналитика тесно связана с машинным обучением. На самом деле, системы ML часто предоставляют инструменты для аналитики интеллектуального программного обеспечения.

На заре появления Big Data компании исследовали свои данные, чтобы понять, что было в прошлом. После этого они начали использовать свои инструменты для анализа, чтобы определять причины тех или иных событий.

Прогностическая аналитика идет еще дальше. Она предсказывает, что произойдет в будущем, используя анализ Big Data. Число организаций, использующих предсказательную аналитику в 2017 году, – не очень большое, всего 29 процентов, согласно опросу 2016 года от PwC.

Тем не менее многие поставщики готовых решений представляют интеллектуальные инструменты для аналитики. И за счет их клиентов количество компаний, использующих предсказательную аналитику, может резко увеличиться.

Большая часть финансовых функций и подразделений также будет заменена алгоритмами и сервисами, позволяющими получать инсайты и ответы на регулярные вопросы со стороны владельцев бизнес-процессов о состоянии дел.

Поменяются и форматы представления данных – в сторону стандартных нотаций (например, XBRL).

Сайты компаний будут иметь интерфейсы для аналитических сервисов, которые будут позволять автоматизировать, например,

отчетность для инвесторов.

Big Data приложения – появляется простота и стабильность

Машинное обучение и технологии ИИ используются для создания приложений. Они, например, анализируют предыдущие действия пользователя, и за счет этого делают персонализированные предложения. Одним из известных примеров являются рекомендательные сервисы, которые сейчас используются множеством приложений для электронной коммерции и развлечений.

Развивается направление Intelligent Security

Многие компании также включают Big Data-аналитику в свою стратегию безопасности. Данные из логов организаций предоставляют информацию о прошлых попытках атак. Их можно использовать для прогнозирования и предотвращения будущих атак.

В результате, некоторые компании интегрируют свое ПО для обеспечения безопасности и управления событиями с платформами Big Data, такими как Hadoop. Другие – обращаются к поставщикам решений по безопасности, чьи продукты включают в себя большие возможности для анализа данных.

Все больше решений IoT

Интернет Вещей тоже вносит вклад в Большие данные. Согласно отчету **IDC**[151],«31,4 процента опрошенных организаций запустили решения IoT, а 43 процента планируют развернуть их в ближайшие 12 месяцев». Со всеми этими новыми устройствами и приложениями, которые появляются в сети, данных будет еще больше, чем раньше. Многим компаниям потребуются новые технологии и системы для обработки возрастающего потока данных, поступающих из их решений IoT. Большую интеграцию и развитие также получат смежные сервисы, где данные с устройств будут использоваться для предоставления сторонних сервисов, например финансовых, таких как страхование имущества или кредитование под поставку объектов имущества.

Развиваются решения Edge Computing

Одной из новых технологий, которые могут помочь

компаниям справиться с Большими данными IoT, являются вычисления на узлах (машинах), близких к источникам данных.

Это называется Edge Computing (англ. edge – «край»). В Edge Computing Big Data-анализ происходит очень близко к устройствам и датчикам IoT, а не в центре обработки данных или облаке. Компаниям это дает существенные преимущества. У них становится меньше данных, передающихся по их сетям. В результате, можно повысить производительность и сэкономить на стоимости облачных вычислений в сети. Это позволяет организациям удалять данные IoT, которые являются ценными в течение ограниченного периода времени, что снижает затраты на хранение и инфраструктуру. Edge Computing также может ускорить процесс анализа, снижая time-to-market для аналитики.

Возрастает ценность людей

Для IT-специалистов рост Big Data-аналитики, вероятно, будет означать высокий спрос и высокие зарплаты для тех, кто смог быстро набрать опыт по работе с Big Data-технологиями. По данным IDC: «Только в США в 2018 году будет 181000 вакансий, связанных с аналитикой, и в пять раз больше позиций, требующих соответствующих навыков управления и интерпретации данных».

Появился целый новый рынок труда со множеством профессий и специализаций, не имеющий пока четких критериев для отбора и поиска специалистов, за исключением рейтингов Kaggle или участием в тех или иных исследовательских проектах. Большим риском в предстоящем развитии новых профессий по работе с данными является в том числе их оторванность от понимания бизнес-специфики, так что специалисты, которые будут совмещать в себе понимание как IT, так и бизнес-составляющей, будут получать высокие зарплаты.

Существенную роль сыграет постепенное появление CDO (Chief Data Officer) в команде руководства большинства компаний. Если проанализировать публичные профили известных CDO, то большинство из них сегодня пришло к этой роли из бизнеса через трансформацию своей компании, сохранив при этом определенный уровень компенсаций и ожиданий.

Растет популярность Self-Service

Поскольку стоимость найма экспертов возрастает, многие организации будут искать инструменты, которые позволят

обычным бизнес-пользователям удовлетворять свои потребности в аналитике данных. Ранее IDC предсказывал, что «инструменты для визуальной работы с данными будут расти в два с половиной раза быстрее, чем рынок бизнес-аналитики (BI). К 2018 году инвестиции в этот инструмент Self-Service конечных пользователей станут обязательными для всех предприятий». Несколько поставщиков уже запустили инструменты для аналитики Больших данных с такими возможностями. Эксперты ожидают, что тенденция продолжится и дальше. IT, скорее всего, будет менее вовлечен в процесс, так как большая аналитика данных относится, в первую очередь, к предметной области, которой занимаются бизнес-пользователи.

Рост объемов данных продолжится

Сегодня компаниям нужно все больше знать о своих продуктах и пользователях и, как следствие, успевать адаптироваться к изменяющимся требованиям со стороны рынка.

Даже промышленный сектор стал активно переходить в область использования аналитики и работы данными. Так, промышленная компания по разработке программного обеспечения Uptake быстро достигла капитализации в один миллиард долларов, получив звание единорога. Суть ее предложения – помогать промышленным компаниям оптимизировать свой бизнес и продукты на основе инсайтов, полученных из анализа при работе с промышленными данными. Компании удалось построить решения для различных индустрий, начиная от транспорта и добычи, заканчивая использованием аналитики для альтернативных источников энергии (ветер и так далее).

Согласно **исследованиям**[152] рост данных для аналитики в реальном времени составит около тридцати процентов в ближайшие два года.

Работа с большими объемами и потоками данных – больше не прерогатива крупных компаний с большими бюджетами, теперь она доступна и среднему, и малому бизнесу. Это стало результатом популярности (и, как следствие, появлению простых упакованных решений) технологий Big Data и уменьшению их стоимости.

In-memory решения

Одна из технологий, которую компании исследуют и начинают применять в попытках ускорить обработку Больших данных, – это in-memory решения. В традиционных БД данные хранятся в системах хранения, оборудованных жесткими дисками или твердотельными накопителями (SSD). In-memory технология хранит данные в ОЗУ, а это во много раз быстрее. В отчете Forrester **Research**[153] говорится, что рост количества данных в in-memory решениях будет составлять 29, 2 процента в год.

Конец Big Data

Термин Big Data постепенно отмирает. Он охватывает слишком много тем.

Развивается и специализация. Скоро говорить: «Я работаю в Big Data» будет так же странно, как и «Я работаю с компьютером». Уже сейчас существует множество дисциплин – от машинного обучения, сбора и управления данными до их безопасности. Эти дисциплины имеют между собой мало общего или вообще не связаны, но все равно относятся к Big Data. Кроме того, Big Data сейчас проникает абсолютно во все сферы жизни, и выделять ее в отдельную отрасль становится бессмысленным. Промышленность, IT, образование и даже дизайн сейчас используют или начинают использовать инструменты Big Data для сбора и анализа данных, появляющихся в процессе цифровизации.

Послесловие

Сегодня данные стали (или становятся) важной частью нашей жизни. Сервисы и продукты становятся цифровыми.

Надеюсь, что эта книга помогла составить общее понимание о том, как работают системы Больших данных и для чего они вообще применяются.

Появляются новые инструменты и фреймворки, которые позволяют работать с данными максимально широкому кругу людей. И поэтому очень важно, чтобы все эти люди говорили на одном языке и хотя бы примерно представляли, как все это работает.

В этом смысле книга полезна как начинающим, так и уже сложившимся специалистам. Она будет интересна тем, кто задумывается о смене карьеры, и тем, кого своя карьера устраивает/кому просто любопытно.

Мир меняется, и сейчас навык анализа данных требуется и юристам, и маркетологам, и множеству других профессий. Во многих организациях сейчас идут кампании по продвижению data-driven культуры, но тут часто дело ограничивается только технической стороной – базовым обучением программированию, SQL и, может быть, вебинарами «Learning для чайников».

Но этого недостаточно. Золотой принцип аналитики – это «Garbage in – garbage out»[154], что означает: никакие технические навыки не заменят умения понимать, откуда данные взялись, насколько им можно доверять и каковы границы их применимости.

Высокоуровневое представление о Big Data важно и для бизнеса. Сотрудники компаний, собирающиеся монетизировать свои потоки данных, могут с ее помощью оценить, насколько их подход к вопросу системный. Те, кто еще этого не делает – оценить, что им (возможно) предстоит сделать в будущем.

«Взгляд с высоты птичьего полета» нужен и обычным людям, никак, казалось бы, не связанным профессионально с миром Big Data. По аналогии с компьютерной грамотностью людям сейчас нужна и data-грамотность. Любой человек сейчас должен понимать, какие «следы из данных» он оставляет, и что с этими данными будет дальше.

Данные, которые мы сейчас довольно бездумно и беззаботно оставляем в публичном доступе, могут остаться там на всю нашу жизнь – и влиять на нее. Яркий пример – расторжение контракта с Джеймсом Ганном из-за твитов, сделанных в 2011 году.

Аналогичная история с данными, которые мы отдаем разным

коммерческим и некоммерческим организациям. Многие ли из нас хотя бы просматривают соглашение об использовании данных при регистрации в новом сервисе? Понятно, что почти никто.

Как эти данные будут применяться, сколько лет они будут храниться, могут ли их кому-то перепродать? Будете ли вы рады, если информация о ваших покупках войдет в данные для скоринговой модели микрофинансовой организации?

Не хотелось бы заканчивать книгу на мрачной ноте. Работа с данными – это увлекательное занятие, результаты которого действительно меняют мир.

notes

Примечания

1

По некоторым оценкам используется цифра 760,6 мегабайт для XX-хромосом и 735,9 мегабайт для XY-хромосом, или используется оценка в 400 мегабайт на один сперматозоид, что, в принципе, еще больше.

2

3

Горелов И. Н., Седов К. Ф. Основы психолингвистики. М., 2001. С. 105–106. Тер-Минасова С. Г. Язык и межкультурная коммуникация. М., 2000. С. 29–30.

4

Горелов И. Н., Седов К. Ф. Основы психолингвистики. М., 2001. С. 105–106. Тер-Минасова С. Г. Язык и межкультурная коммуникация. М., 2000. С. 29–30.

5

Ханс Геста Рослинг – шведский врач, академик, профессор Каролинского института по вопросам международного здравоохранения, специалист по статистике и всемирно известный лектор.

6

E-Gov – технологично-центрированная, реактивная среда предоставления государственных сервисов в электронном формате. Начальный этап развития цифрового государства, который измеряется процентом покрываемых существующих сервисов в электронном виде.

7

В соответствии с 152-ФЗ «О персональных данных».

8

В соответствии с 152-ФЗ «О персональных данных».

9

В соответствии с письмом ФНС РФ от 23.11.15 № 11–06/0733, поле «ИНН» стало обязательным реквизитом при подаче справки 2-НДФЛ.

Позднее размер штрафа предлагалось увеличить до пятисот рублей с одной записи, где нет обязательного атрибута ИНН.

11

Эти главы я писал под действием сильных психотропных препаратов, поэтому они могут показаться вам глубокими и сложными.

Но без них практически невозможно понять, о чем здесь написано.

12

13

What data for data-driven learning? Alex Boulton, 2011 Nottingham. https://files.eric.ed.gov/fulltext/ED544438.pdf

14

Согласно Wikipedia, ко́рпус (в данном значении множественное число – ко́рпусы, не корпуса́) – подобранная и обработанная по определенным правилам совокупность текстов, используемых в качестве базы для исследования языка.

15

Согласно Wikipedia, это – статистическая модель, имитирующая работу процесса, похожего на марковский процесс с неизвестными параметрами, и задачей ставится разгадывание неизвестных параметров на основе наблюдаемых. Полученные параметры могут быть использованы в дальнейшем анализе, например, для распознавания образов.

16

Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?

17

18

Книга «Черный Лебедь. Под знаком непредсказуемости» Насим Таллеб.

19

Рекомендую посмотреть документальный фильм – АльфаГо.

20

Приобретена Google.

21

Оценка известного писателя-фантаста Вернона Винджу – 2030, а известного изобретателя и футуролога Реймонда Курцвейла – 2045.

22

Известный эксперт Малькольм Чисхолм (Malcolm Chrishom), который работает в области управления данными более 25 лет, подготовил и опубликовал концепцию жизненного цикла данных.

23

По версии DAMA Internation – независимая некоммерческая профессиональная организация, разрабатывающая стандарты по управлению данными DMBOK (Data Management Book of Knowledge).

Например, по модели Johnson и Scholes.

25

26

По итогам обзора Digital Rights Center.

27

28

Американский ученый австрийского происхождения; экономист, публицист, педагог, один из самых влиятельных теоретиков менеджмента XX века.

29

По оценке Emeritas, одного из ведущих американских агентств по исследованию данных и расчета жизненного цикла для клиента (customer lifetime value).

30

Товар или услуга, которые предоставляются различными конкурирующими компаниями, но качество товара или услуги при этом никак не меняется.

31

Data Silos – или резервуар данных, это фиксированные данные, которые находятся под контролем одного департамента и не передаются другим департаментам.

Garded Goh, Status и другие.

33

По версии Harward Business Review.

Организована при поддержке Stanford Center on Philanthropy and Civil Society (Stanford PACS, Bill & Melinda Gates Foundation, трейдинговой компанией Liquidnet, и Knight Foundation).

36

По версии Medium.com.

Например, создания единой канонической модели данных для передачи по интеграционному слою для всех источников данных внутри организации.

39

В статье Laurel Brulk, эксперт в области данных и маркетинга, указывает на особенности профессии data engineer.

41

Непрерывная фрактальная заполняющая пространство кривая, являющаяся вариантом кривой Гильберта.

Организована при поддержке Stanford Center on Philanthropy and Civil Society (Stanford PACS, Bill & Melinda Gates Foundation, трейдинговой компанией Liquidnet и Knight Foundation).

43

По оценке в конфигурации из 16 ТБ

45

Согласно **Definition of Done**, в полной мере понятна людям знакомыми с философией Scrum. Под определение сделанной задачи попадает задача, которая не нуждается в доработках.

50

53

Известный консультант в области данных, который проработал в различных компаниях таких как Adobe, Test&Target и других.

Антонио Дамассио.

55

60

61

Взаимосвязанные этапы единого процесса по привлечению новых клиентов и продаже продуктов.

62

См. книгу SAS Best Practices “Storytelling in Business” by Bree Baich и Analise Polsky. https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/storytelling-n-business-109014.pdf

63

Было позднее в 2007 году приобретено Google Inc и позднее решение было переименовано в Google Public Data Explorer. В 2016 оно включило в себя так же все возможности Google Analytics Suite.

65

71

Выступление Джефри The Future of Data Visualisation 2015 – Strata + Hadoop World Conference (San Jose).

77

78

80

Репозиторий с программой позволяющей обрабатывать PDF-файлы, которые публиковали власти города Нью-Йорк.

81

82

Бил де Блазио, мэр города Нью-Йорка с 2014 года.

Health Insurance Portability and Accountability Act (HIPAA).

88

89

92

94

Flash Boys: Высокочастотная революция на Уолл-Стрит. Автор: Майл Льюис, 2014 год.

95

100

101

102

103

106

107

108

109

110

111

112

113

115

116

117

119

Международные стандарты финансовой отчетности.

В разных аудиторских практиках есть разные подходы. Некоторые из них определяют, что необходимо контролировать размер всех неучтенных разниц ниже «порога чувствительности», чтобы они в своей сумме не превышали сам размер порога. То есть, если порог чувствительности 1 млрд руб., то сумма всех разниц ниже этого порога, должна быть меньше 1 млрд руб., тогда аудитор не обращает на это внимания.

121

Для тех, кто находится в глухом и недосягаемом танке, имеется в виду специальное мобильное приложение, через которое люди могут видеть состояние своих банковских счетов и выполнять банковские операции.

122

Организация, которая в соответствии с законом обязана платить налоги на получаемый доход в ФНС.

123

Справка о доходах физического лица, полученных за определенный отчетный период в конкретной организации.

124

Ставка, под которую Центральный банк всем кредитным организациям выдает деньги.

125

127

Вначале 2000-х «Горбушкой» назывался рынок, находящийся рядом со станцией метро «Багратионовская» в Москве, где можно было купить на отдельном компакт-диске (лазерный диск) всю необходимую базу данных о различных пользователях.

128

Веб-версия приложения для работы с банковскими операциями, куда можно попасть по специальному логину.

129

Имеется в виду разделение на тех, кто зарабатывает деньги (то есть заключает сделки и продает сервис), и тех, кто поддерживает работу организации.

Так я ласково называют всевозможные IT-сервисы, связанные с данными, то есть просто тупо обобщаю, чтобы даже простые смертные смогли это понять, как настоятельно просило меня издательство.

131

Правление минус две позиции, то есть начальник управления или подразделения.

132

133

Файлы, в которые записываются все события, происходящие в каждой системе.

http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf

135

137

20 % усилий дают 80 % результата.

138

Master Data Management – платформа управления мастер-данными.

Например, выпадающие списки, где можно выбрать значение только из списка.

Так как источников ввода может быть много, то допускается, что они сгружают из систем в решение по тому или иному домену по обработке данных.

Централизованным он называется потому, что инструмент становится мастер источником для всех остальных систем и распространяет заведомо качественный и согласованный контент, который в нем появляется.

Customer Data Integration platform.

Reference Data Management.

Нормативно-справочная информация.

145

Product Information Management.

Product Lifecycle Management.

147

Согласно одному из определений, данных в журнале Harvard Business Review.

148

Event-driven process chain – событийная цепочка процессов, определенный тип блок-схемы, используемый для бизнес моделирования. Существуют различные вариации использования нотации (alternative, extended и так далее).

149

151

(англ.) «Мусор на входе – мусор на выходе».